**Project Title: Insight into the Analyst Job Market**

|  |  |
| --- | --- |
|  | **Ruth Mulu** |

# Dataset Description

* 1. **Introduction**

While taking on this project, I contemplated a couple of ideas where I can explore my data gathering, cleaning, and analysis skills. I wanted to pick a research topic that I can take as an input in my future decision making and give me a meaningful insight on the area I would embark on. So, I chose to analyze the data of current and past job market trends of data analyst job positions in the United States.

Embarking on this analysis would help me understand:

* Labor market trends
* Salary structures and it’s dependencies
* Skills needed
  1. **Data Source and Findings**

I tried to get data from the United States Department of Labor, but the dataset was not quite intuitive and up to date. So since almost all data analyst job positions are handled by online websites, I opted to use data from online sources. After careful evaluation I opted to use open resource data that is compiled by Kaggle, which is Google’s data science online community for data scientists and machine learning practitioners.

For this task, I imported two sets of data, each having different attributes. The target is to combine the data for better cleansing, analysis and insight with the aim of having a better understanding of the past trends, current interests and prospects.

**Dataset 1: Data Analyst Job from Glassdoor**

This is a dataset that shows the job market landscape after the pandemic. I choose to use this dataset because I wanted to understand the trend of the job market back in June 2020. This dataset is specifically filtered to show the status for only the positions posted with “Data Analyst” description in different industries and locations.

Source URL: <https://www.kaggle.com/datasets/andrewmvd/data-analyst-jobs/data>

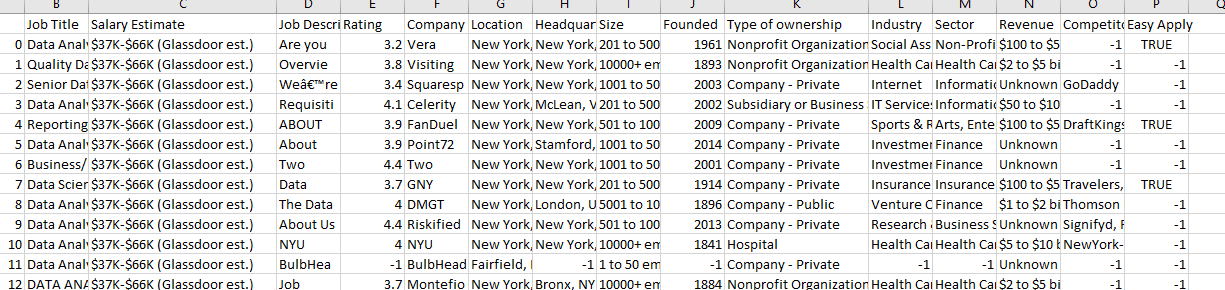


Fig- Uncleaned dataset 1

**Dataset 2: LinkedIn Job Postings (2023 - 2024)**

This dataset contains a nearly comprehensive record of 123,000+ job postings listed in 2023 and 2024. Each individual posting contains dozens of valuable attributes for both postings and companies. As LinkedIn is now one of the most popular websites for networking and job search, this data will surely give me insight on the demand. This data will help me explore the domain beyond the niche data analyst positions and include related positions like financial analyst and Business analyst.

Source URL: <https://www.kaggle.com/datasets/arshkon/linkedin-job-postings>

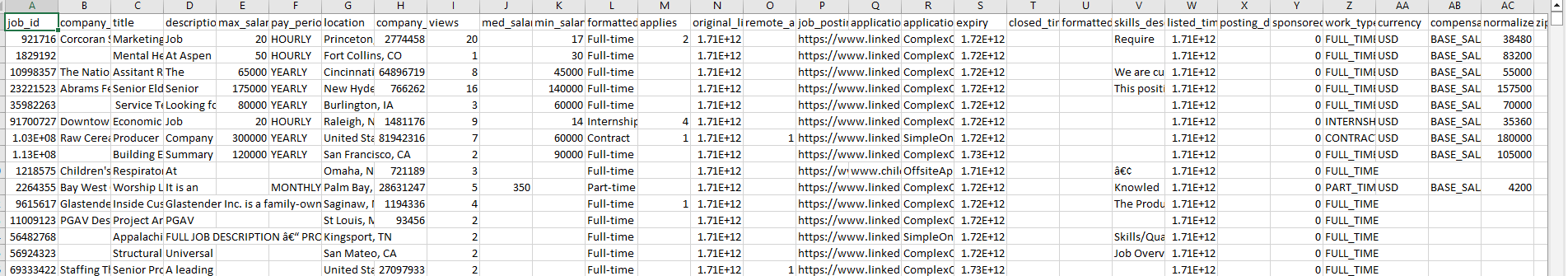


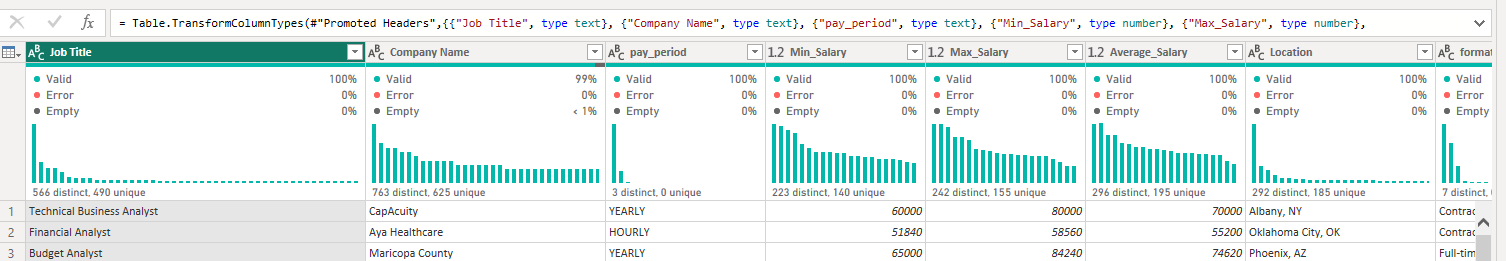
Fig- Uncleaned dataset 2

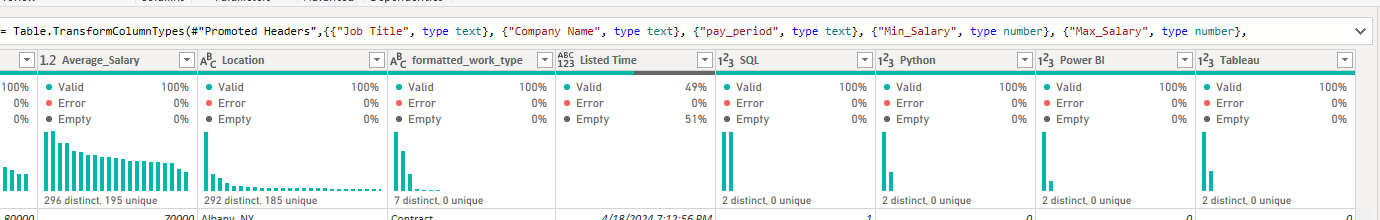
* 1. **Data Source Comparisons:**

|  |  |  |
| --- | --- | --- |
| **Data Name** | **Data Analyst Job from Glassdoor** | **LinkedIn Job Postings** |
| Source | Kaggle | Kaggle |
| Data Source Country | USA | USA |
| Primary data collection method | Collected by Glassdoor; then made Open source available | Collected by LinkedIn; then made Open source available |
| File Format | .csv | .csv |
| Data Frequency/ quality | *(Shown below)* | *(Shown below)* |
| Real-time | No | No |
| Original Data Period | 2020 | 2023 – 2024 |
| Is the dataset Structured | Yes | Yes |
| Data collected from | Glassdoor Jobs | LinkedIn Jobs |
| Job | Only Data Analyst | All job postings |
| Data Size (Rows) | 2,253 | 123,842 |
| Attributes/ Column size | 15 | 31 |
| Variable coverage or Important Column Attributes | Job Title; Salary Estimate; Job Description; Company Name; Location; Size; Industry; Sector; Revenue; Competitors; Easy Apply | Company Name, Job title; Job Description; Max salary; Location; Med salary; Min salary; Skills Desc; |
| Data Pros | * Data concentrated on Data Analyst * 100% data quality on salary * Industry specific | * Data is very recent * Data size is large * Data specified the posting and expiry dates |
| Data Cons | * Job description is not categorized * Data lacks to specify date of posting * Data size is small * Data is not recent | * Job description is not categorized * A lot of unwanted data that needs filtration * Less data quality on salary * Industry not specified |

As both datasets offer a different attribute on their own, combining them would offer a better tuning to my analysis.

Data Quality and distribution lookup using Power BI:





* 1. **Software/ Tool Used**

1. Microsoft Excell: Mostly for data cleaning and conversions
2. SAP Analytics: For regression analysis
3. SAP Cloud: for visualization
4. Power BI: For data quality, distribution and profiling. It was also used for random crosschecking of results.

# Data Analysis Methodology

## Data Analysis Techniques

Although there are a lot of interdependencies on job applications, prior experience, skills, locations, and salaries among others; I tried to summarize roughly the scenery of the market using the available dimensions only.

For my case, I would use the following techniques for a better analytical understanding:

1. **Descriptive analysis**:

I used this technique to have a holistic picture about the dataset and also get an insight on a specific attribute. This technique is used here to better understand common trends and summarize the data based on my chosen dimensions.

1. **Regression Analysis**:

To understand if there is any relationship between specific skill and salary estimation, I used regression analysis to understand if there is any correlation between the two.

* 1. **Data Cleaning Approach**

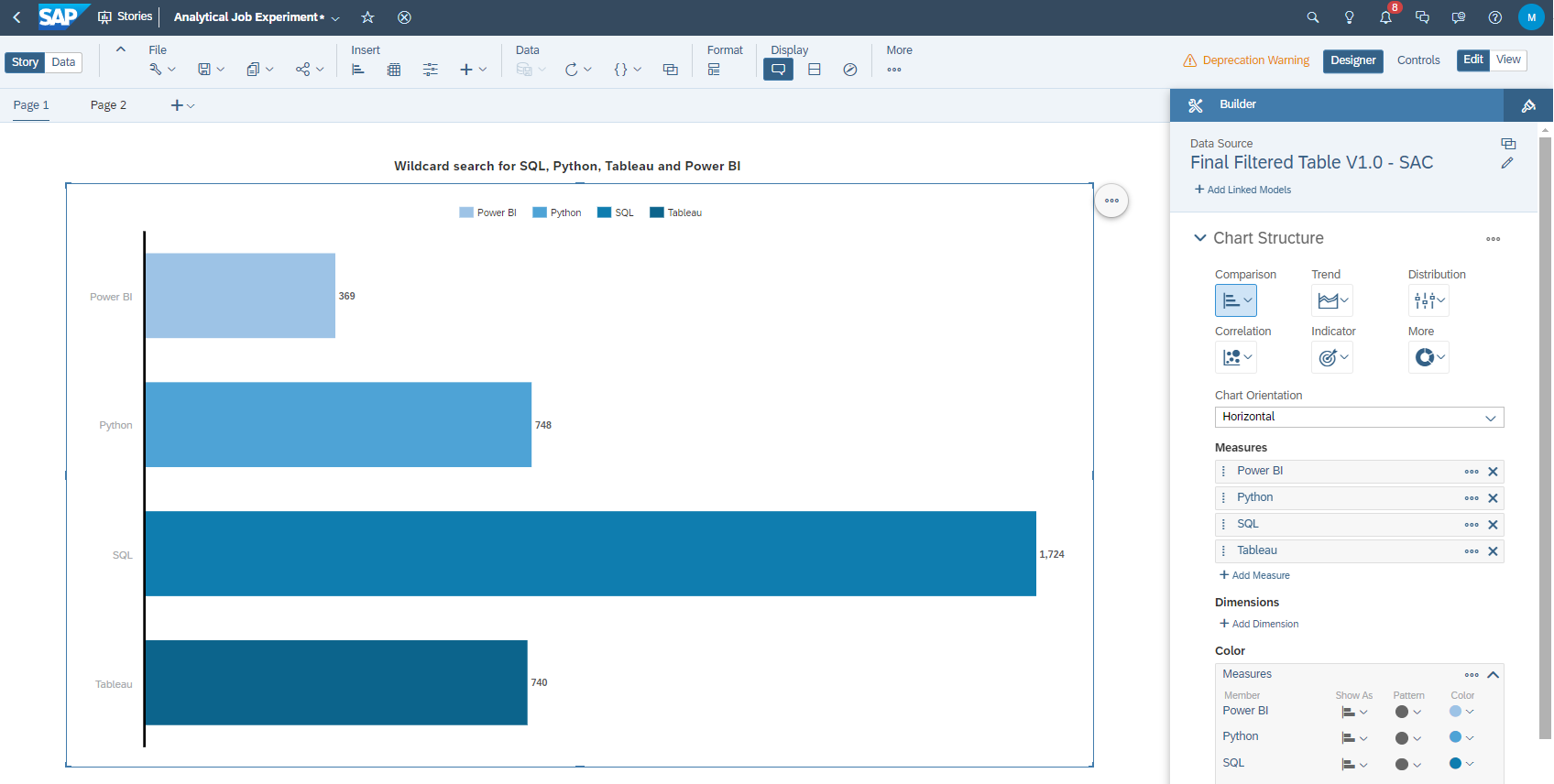
As both datasets follow a different format, I have done the following cleansing actions in order to synchronize the data formatting and get it ready for analysis. While most of the cleaning is done on excel, some have been done on SAP Predictive Analytics.

|  |  |
| --- | --- |
| **Applied Dataset** | **Actions Taken** |
| Both datasets | Data filtration (*Mostly applied LinkedIn Job Postings as it contains a lot of job descriptions that are out of the topics outlined here*) |
| New calculated columns formed including Average salary |
| Data type conversion applied on the numbers and dates |
| Advanced excel filtering to filter the job titles that only include the below wildcards:  \*Data Analyst\*  \*Data Analysis\*  \*Business Analyst\*  \*Business Analysis\*  \*Financial Analyst\*  \*Financial Analysis\*  \*Business Intelligence Analyst\*  \*Business Intelligence\*  \*Operation Analyst\*  \*Budget Analyst\*  \*System Analyst\*  \*Report Analyst\*  \*Functional Analyst\*  \*Marketing Analyst\*  \*Market Analyst\*  **Result: 3400 Jobs filtered using Microsoft Excel’s advanced wildcard filtration from 126,102 source Job postings.** |
| Wild card lookup was used to search for the below items from the job description:  **SQL, Python, Power BI and Tableau**  Function example: *=IF(ISNUMBER(SEARCH("\*SQL\*",C2)), 1,0)* |
| Glassdoor Jobs | Removal of unnecessary content from the table to find matching relevant dimensions including job ratings, company establishment period, competitions etc. |
| Using delimiter functions to remove symbols and convert salary to Minimum and Maximum salary columns |
| The skill description and the job description have been merged for a better localized searching |
| LinkedIn Jobs | Removal of unnecessary content from the table to find matching relevant dimensions including job id, job link, expiry date, application link etc. |
| Hourly, weekly, Bi-weekly and monthly ratio conversion by using the below formula:  *=IF(F2="HOURLY",D2\*12\*4\*40,IF(F2="WEEKLY",D2\*12\*4, IF(F2="BIWEEKLY",D2\*12\*2,IF(F2="MONTHLY",D2\*12,IF(F2="YEARLY",D2,D2)))))* |
| Format Conversion: Converting the 13 digit date format to a correct date with function: *=(N2/86400000)+DATE(1970,1,1)* |

# Analysis and Results

1. **The number of times a skill appeared on a wildcard search**

Excluding the major analytical and visualization tool, Excell, I took into account two analytical tools/languages (SQL and Python) and two visualization software (Power BI and Tableau); and then tried to see how many times they appeared on the job description and skill requirement sections.

****

**Fig 1: Wildcard search for SQL, Python, Tableau and Power BI**

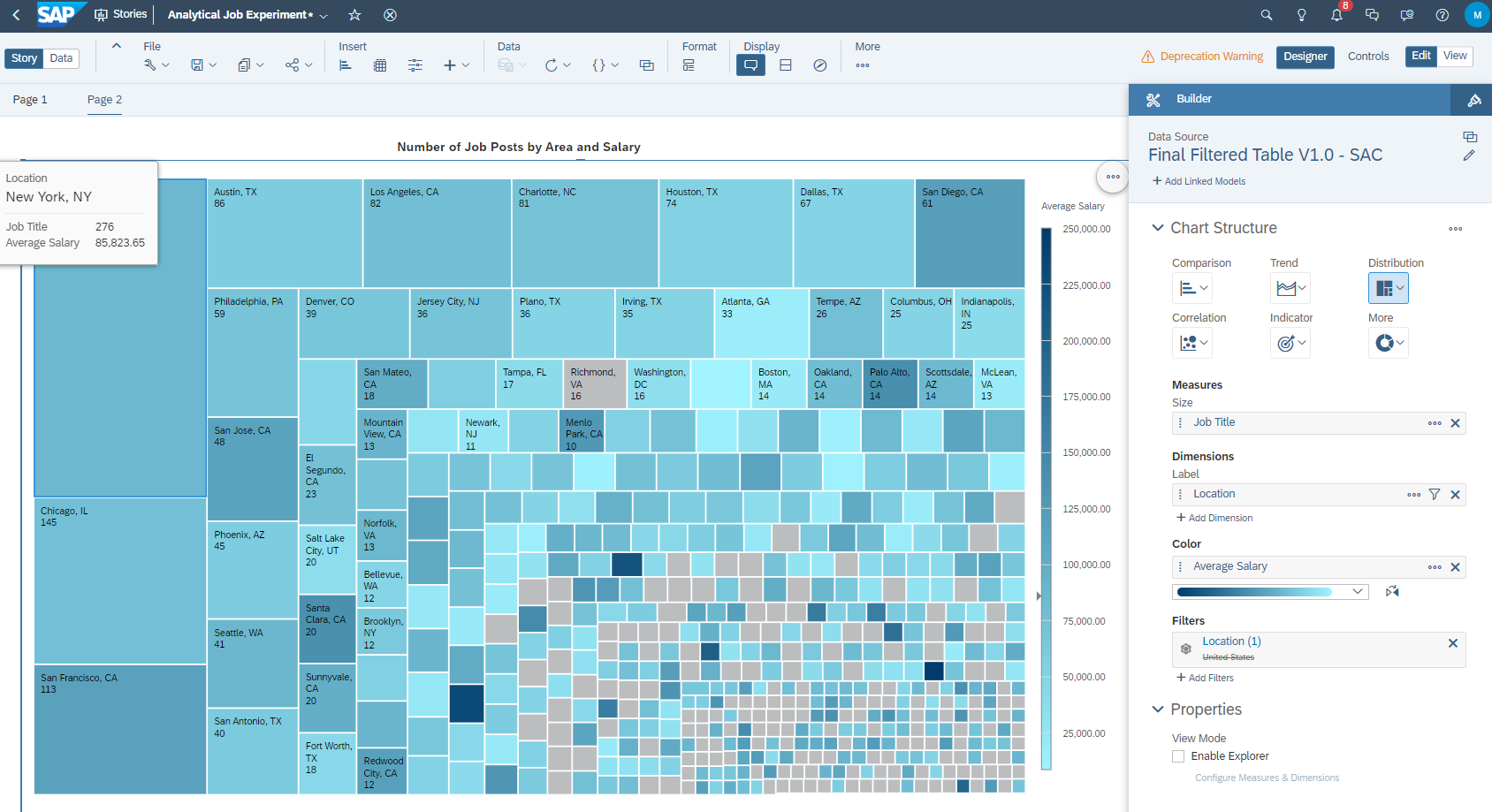
From the above skills needed on the filtered job titles, SQL appeared the most. Here is the list:

1. **SQL** (1724 Mentions)
2. **Python** (748 Mentions)
3. **Tableau** (740 Mentions)
4. **Power BI** (369 Mentions)
5. **Areas with the most job search posts**

I incorporated the location of the company that posted the job application and tried to summarize by area where there are more job vacancies and their respective salaries.

From the analysis result, here are the top 3 areas with the most job post along with their salary

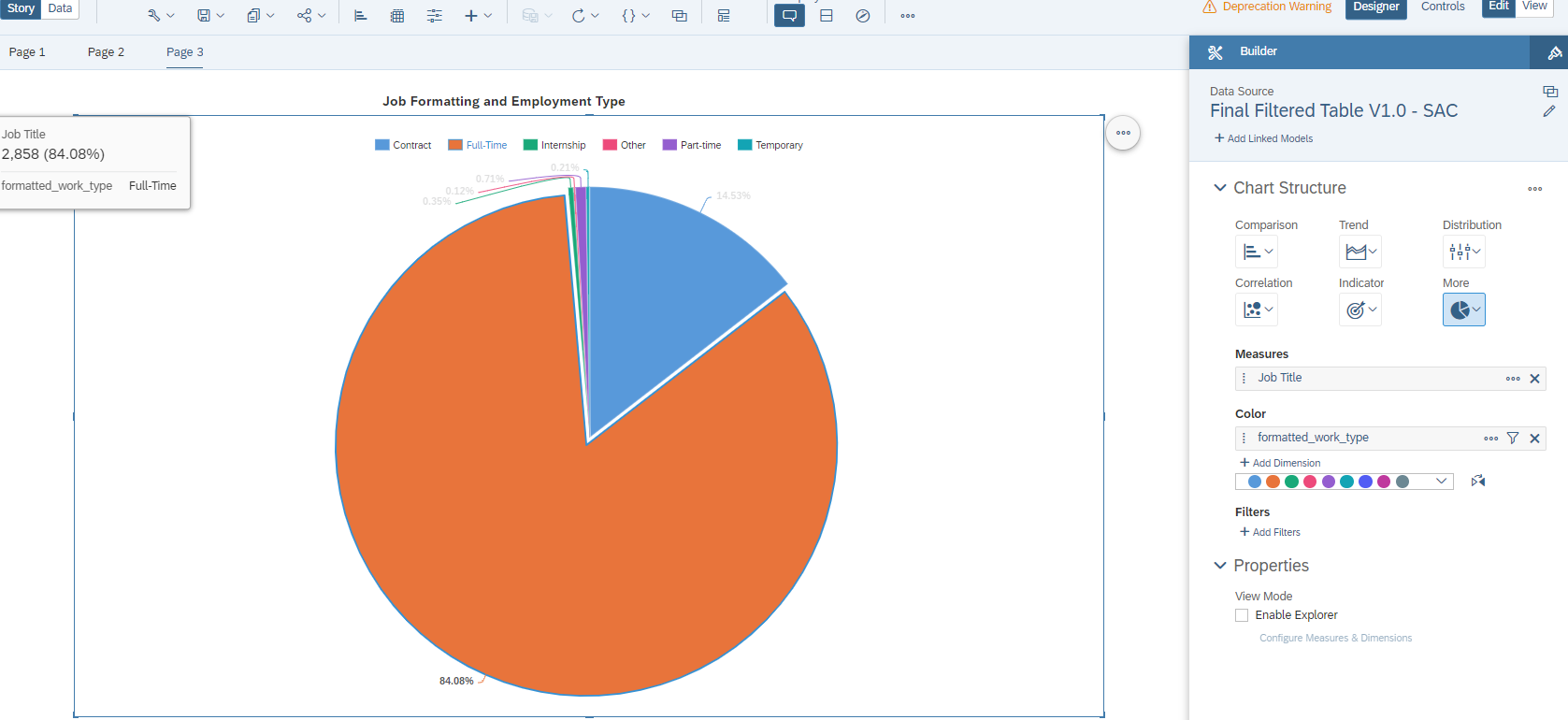
1. **New York, NY**: with 276 relevant Job Post (Avg. salary 74.5K)
2. **Chicago, IL**: with 145 relevant Job Post (Avg. salary 79.9K)
3. **San Francisco, CA**: with 113 relevant Job Post (Avg. salary 100.2K)



**Fig 2: Job Vacancies by area and salary**

1. **Job Formatting/ Employment Type**

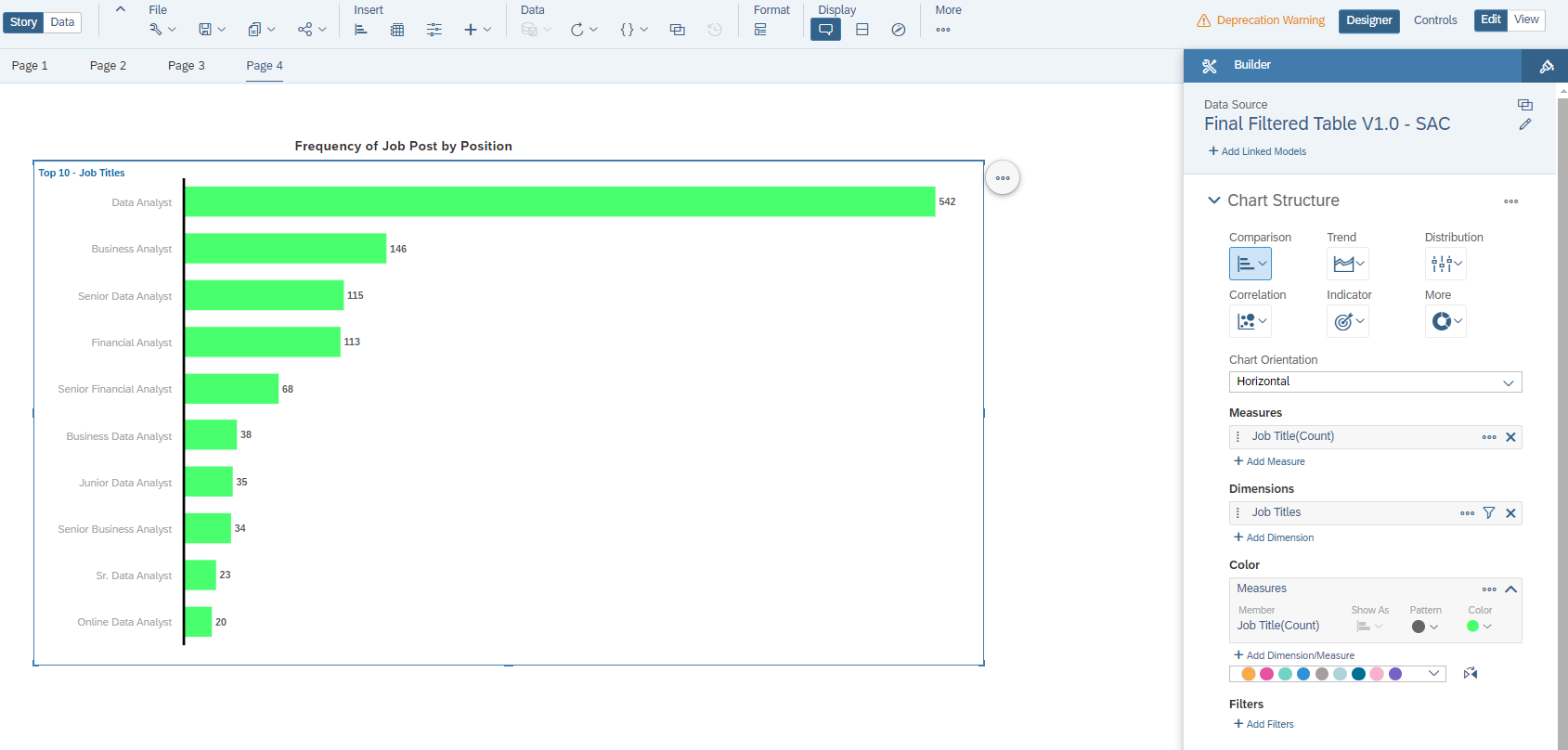
Furthermore, I wanted to check how many portions of the jobs that are being posted are Full-time and how much of those were contractual and internships. Below is the pie chart that shows that **84% of the job vacancies are full-time**.



**Fig 3: Job vacancies by employment type**

1. **Top 10 Frequent Position Vacancies**

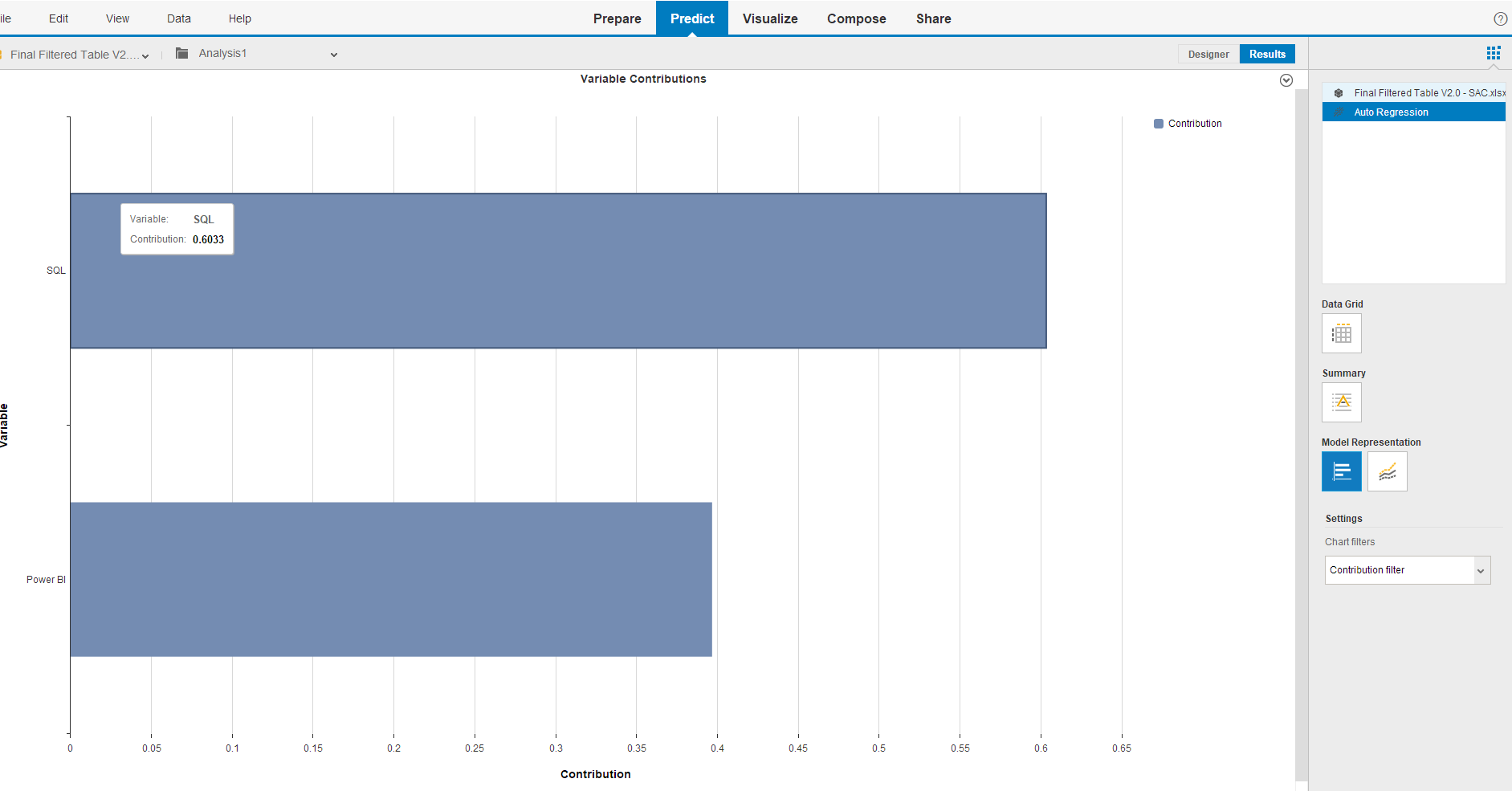
From the vacancies that are made available, I looked at the 10 roles by frequency. **Data Analyst**, **Business Analyst** and **Senior Data Analyst** are consecutively the most frequently posted vacancies.

****

**Fig 4: Top 10 Frequently posted job positions**

1. **Major Skill Contributors for Salary using regression analysis**

I utilized SAP Analytics’ Auto regression tool to check the top skill influencers of high salary. Each row has a record of the existence or absence of these skills in their description. The system was able to find out that **Power BI and SQL are the most contributors to high salary**.

****

**Fig 5: Top Skill Contributors to high salary**

**Note**: The outputs of the data were exported and attached separately for reference purposes. The pictures are clearer on the formatted pdf document. Due to exporting limitations, the regression can only be viewed but couldn’t be exported.



# Impact of the Analysis

Analytics plays a pivotal role in shaping decision-making processes across industries. By analyzing vast amounts of data, anyone can uncover hidden patterns, trends, and relationships to drive a certain decision. Specifically, in the job market, analytics helps businesses and job seekers alike to understand labor trends, salary structures, and the demand for specific skills. If the right amount and quality of data is available, through the different tools at my disposal, it has become easier to forecast industry trends, and better align strategies with their market demands.

For example, in the context of the data analyst job market, analytics enables companies to determine the impact of technical skills such as SQL, Python, Power BI, and Tableau on salaries and job openings. By leveraging data from platforms like LinkedIn and Glassdoor, job seekers can also gain insights into which skills are most valued, where job opportunities are concentrated, and how different employment types of influence salary offers.

# Reflection on the overall analysis

In terms of tool utilization, Since the dataset I used was very large, I was able to test the efficiency of four different platforms (excel, SAP Analytics and SAP Cloud and Power BI). Due to my prior experience, I did most of the cleaning and analytical work with Microsoft Excel. On the contrary, I found out that the efficiency was very low as it took me a lot of time to process the data. When you are using excel for a large dataset, it would also require a computer with a excellent processing speed.

By analyzing the above data points, I was able to observe that skills such as SQL are highly sought after, while tools like Power BI are associated with higher salary packages. Furthermore, the geographical analysis highlighted those metropolitan areas like New York and San Francisco offer higher salaries, reflecting the concentration of tech companies and demand in those regions.