**Business Analyst Dataset**

**Overview**

After checking the dataset for errors and duplicates it now contains 4035 rows and 16 columns on Business Analyst Jobs. The tools I used to analyse this dataset was R and RStudio. I conducted data exploration to identify patterns and relationships across the dataset.

I wrote 15 queries in R Studio and performed my analysis by using calculations to gain an insight about the data. I also created plot visualisations using R Studio to support my findings.

**R Packages and libraries used in my Analysis**

* Readxl – This package makes it easy to get data out of Excel and into R. `readxl` has no external dependencies, so it’s easy to install and use on all operating systems. It is designed to work with tabular data.
* Tidyverse – This is a collection of R packages that share a high-level design philosophy and low-level grammar and data structures, so that learning one package makes it easier to learn the next. I installed all the packages in the `tidyverse` by running `install.packages("tidyverse") then run `library(tidyverse)` to load the core `tidyverse` and made it available in my current R session.
* Dplyr – This is a package in R that provides a grammar of data manipulation, with a set of verbs that help to perform common data manipulation operations on data frames. It is part of the wider `tidyverse` in R and is an upgraded version of the `plyr` package. `dplyr` makes data exploration and data transformation easy and fast in R.

Some of the most commonly used `dplyr` verbs include:

- `select()` to select columns from a data frame.

- `filter()` to filter rows based on a condition.

- `arrange()` to arrange rows in ascending or descending order.

- `mutate()` to add new columns or modify existing columns.

- `summarise()` to summarise multiple values into a single value.

* Ggplot2 is a popular data visualization package in R based on the 'Grammar of Graphics'. It provided me with a flexible system for creating graphics, allowing me to build complex plots layer by layer. With `ggplot2`, I created a wide variety of visualizations, from simple scatter plots to more complex visualizations such as heatmaps, boxplots, and histograms.

`ggplot2` is used for creating graphics, making it easy to learn and use. I created basic plots by specifying the data, mapping variables such as position, colour, and size, and adding layers such as points, lines, or bars. I then customized the plots by adding scales, axes, labels, and other elements.

* Plotly is a package in R that allows you to create interactive web-based graphs via the open-source JavaScript graphing library `plotly.js`. With `plotly` I created an interactive 3D scatter plot to map data variables to visual properties using various arguments. I then added layers to the plot and customized it using additional functions.
* Stringr is a package in R that provides a cohesive set of functions designed to make working with strings as easy as possible. `stringr` focuses on the most important and commonly used string manipulation functions.

Some of the most commonly used `stringr` functions include:

- `str\_length()` to compute the length of strings.

- `str\_c()` to concatenate strings.

- `str\_sub()` to extract or replace substrings.

- `str\_detect()` to detect the presence or absence of a pattern in a string.

- `str\_replace()` to replace matched patterns in a string.

**Summaries**

**1. Load the dataset to R Studio:**

I loaded the library (readx1) and imported the clean dataset from excel.

A screenshot of a computer

Description automatically generated with medium confidence

**2. Display the following columns: Job Title, Rating, Location, Industry**

I installed a package called “tidyverse” and loaded a library called (dbplyr).

Then I created a query to show the specific columns only from the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

**3. Show a list of the following:**

**Top 20 Industries (must be unique values, no duplicates)**

A screenshot of a computer

Description automatically generated with medium confidence

This dataset displays the most unique Industries, with Health Care Services and Hospitals being the top industry.

**Top 20 Sectors (must be unique values, no duplicates)**

A screenshot of a computer

Description automatically generated with medium confidence

This dataset displays the most unique Sectors, with Health Care being the top Sector.

**Top 20 Headquarters (must be unique values, no duplicates)**

A screenshot of a computer

Description automatically generated

This dataset displays the most unique headquarters, with New York being the top location.

**4. Show a list of the following:**

**Top 15 Jobs based on Rating**

**A screenshot of a computer

Description automatically generated with low confidence**

According to this dataset, Business Analysts make up 55% of the selection among the 15 highest-rated jobs.

**Top 15 Jobs based on Rating under “Consulting” industry**

**A screenshot of a computer

Description automatically generated with low confidence**

This displays the 15 highest-rated jobs in the consulting industry, with Business Analysts making up almost half of the selection.

**Bottom 15 Jobs based on Rating**

**A screenshot of a computer

Description automatically generated with medium confidence**

According to this dataset, Business Analysts make up over 85% of the selection among the 15 lowest-rated jobs.

**5. Then plot the following:**

**Top 10 Companies with a Rating greater than 3 and under industry “Consulting”**

**A screenshot of a computer

Description automatically generated with low confidence**

**A picture containing text, screenshot, diagram, line

Description automatically generated**

According to this dataset, Crossfire Consulting is the top-rated company in the Consulting Industry, with a rating of 5 out of 5 based on 5 reviews.

**Top 10 Companies with a Rating greater than 3 and under industry “Energy”**

**A picture containing text, screenshot, font, line

Description automatically generated**

**A picture containing text, screenshot, diagram, line

Description automatically generated**

According to this dataset, Sunnova is the top-rated company in the Energy Industry, with a rating of 4.2 out of 5 based on 3 reviews.

**Top 10 Companies with a Rating greater than 3 and under industry “Accounting”**

**A picture containing text, screenshot, font, line

Description automatically generated**

**A picture containing text, screenshot, diagram, line

Description automatically generated**

According to this dataset, Sysintelli is the top-rated company in the Accounting Industry, with a rating of 4.6 out of 5 based on 5 reviews.

**6. Add 10 more summaries/ plots based on your own analysis.**

**Export the plots and then paste the screenshots into the document:**

**Histogram of Companies that were founded after 1700:**

**A picture containing text, font, handwriting, line

Description automatically generated**

**A picture containing screenshot, diagram

Description automatically generated**

According to this dataset, a significant number of companies were established between 1980 and 2010, indicating that they are relatively new and may not have a long history or heritage.

**A Boxplot of Sector and Rating out of 5:**

****

**A picture containing text, diagram, screenshot, number

Description automatically generated**

When I compare the position of the median line to the central tendency I can see three sectors that stand out;

Transportation and Logistics

Information Technology

Business Services.

In addition, only Information Technology and Education Sectors have a median line above the Rating of 4 across all sectors.

**Density Plot of Sectors founded after 1980:**

**A picture containing text, screenshot, font

Description automatically generated**

**A picture containing text, screenshot, diagram, plot

Description automatically generated**

According to this density plot, a significant number of companies in three Sectors were founded in the years between 1985 and 2000:

Accounting & Legal

Oil, Gas, Energy and Utilities

Retail

However I now wanted to produce another plot to look at the frequency of the Sector to see how many companies were setup in this period.

**Scatter Plot of Sectors founded after 1985 but before 2000, by Revenue:**

**A picture containing text, font, screenshot, line

Description automatically generated**

**A picture containing text, screenshot, font, number

Description automatically generated**

According to this scatter plot based on frequency of Sectors, a significant number of companies in three sectors were founded in the years between 1985 and 2000:

Information Technology

Business Services

Finance.

Now I wanted to delve deeper into the Information Technology Sector and see if I can find any interesting insights.

**Pie chart of Information Technology Sector and the Type of Ownership:**

**A close-up of a white background

Description automatically generated with low confidence**

**A picture containing text, screenshot, diagram, colorfulness

Description automatically generated**

According to this pie chart the type of ownership for Information Technology sector is overwhelmingly Private Ownership, above 75%.

**Pie chart of Information Technology Sector and the Revenue:**

**A picture containing text, font, screenshot, line

Description automatically generated**

**A picture containing text, screenshot, colorfulness, diagram

Description automatically generated**

According to this pie chart the revenue for Information Technology sector is varied and category $50-£100 million (USD) accounts for nearly a quarter 25% of all revenue.

**Bar chart of Information Technology Sector and Location where count is greater than 50:**

**A picture containing text, font, handwriting, screenshot

Description automatically generated**

**A picture containing text, diagram, screenshot, plot

Description automatically generated**

According to this bar chart the top location for Information Technology sector is Austin in Texas

I was rather surprised that Texas was higher that California and I decided to investigate this further and found the following:

1. Texas has favourable tax laws and more lenient regulations compared to other states. For example, Texas has no corporate or individual state income tax. This can make it attractive for businesses to relocate to Texas.

2. Texas has a lower cost of living compared to other states with large tech industries, such as California and New York. This can make it easier for companies to attract and retain talent.

3. In addition to these factors, Texas has a growing pool of tech talent and a reputation for being business-friendly. These factors combined can make Texas an attractive location for Information Technology companies.

**Bar chart of Information Technology Sector and Headquarters where count is greater than 20:**

**A picture containing text, font, screenshot

Description automatically generated**

**A picture containing text, screenshot, diagram, plot

Description automatically generated**

According to this bar chart the top Headquarters for Information Technology sector is Woodridge in Illinois.

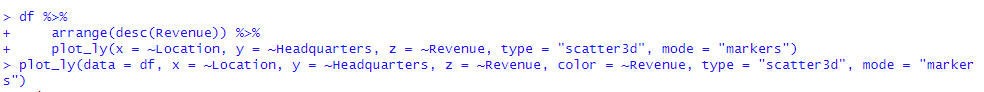
I was rather surprised that again another state such as Illinois was higher that California and I decided to investigate this further and found the following:

1. Woodridge is located in the Illinois Technology and Research Corridor, a region of commerce and industry located along Interstate 88 in the Chicago metropolitan area, primarily in DuPage, Kane, and DeKalb Counties. The corridor is home to the headquarters or regional centres for many Fortune 1000 companies

2. Illinois has offered various incentives to attract and retain tech companies since 2000. Data centre incentives in Illinois have rejuvenated the Chicago-area cloud ecosystem, helping the state attract more than $4.2 billion in new investment

3. Illinois Growth and Innovation Fund (ILGIF), an evergreen, $1 billion impact investment fund established by Illinois Treasurer Michael W. Fredrich’s that invests in ways to attract, assist and retain quality technology-enabled businesses in Illinois.

**3D Scatterplot of Location, Headquarters and Revenue**

****

**A picture containing text, diagram, screenshot, map

Description automatically generated**

According to this 3D scatterplot the revenue for headquarters and location the unknown / Non – Applicable has the highest and followed by $1 to $2 billion (USD) revenue.

**7. Add 25 more summaries based on this dataset.**

**Create a query to show the specific columns only from the dataset.**

df\_subset1 of Job Title, Salary Estimate, Size and Industry:

**A screenshot of a computer

Description automatically generated with medium confidence**

**Now using df\_subset1 create a summary of Top 15 Salary Estimate sorted in descending order**

**A screenshot of a computer

Description automatically generated with medium confidence**

This dataset displays the Top 15 Salary Estimates with all results having a result of $99K -$118K Salary Estimate.

**Now using df\_subset1 create a summary of Top 15 Salary Estimate (must be unique values, no duplicates)**

**A screenshot of a computer code

Description automatically generated with medium confidence**

This dataset displays the most unique Salary Estimates for the Top 15 results.

**Top 15 Jobs based on Salary Estimate under “Health Care Services & Hospitals” Industry**

**A screenshot of a computer

Description automatically generated with medium confidence**

This dataset displays the Top 15 Jobs under “Health Care Services & Hospitals” Industryhaving a result of $99K -$118K Salary Estimate to $72K – $134K.

**Lets look at some other industries and compare their Salary Estimates**

**A screenshot of a computer

Description automatically generated with low confidence**

**Top 15 Jobs based on Salary Estimate under “Biotech & Pharmaceuticals” Industry**

**A screenshot of a computer

Description automatically generated with medium confidence**

This dataset displays the Top 15 Jobs under “Biotech & Pharmaceuticals” Industryhaving a result of $94K -$163K Salary Estimate to $69K – $97K.

**Top 15 Jobs based on Salary Estimate under “Venture Capital & Private Equity” Industry**

**A screen shot of a computer code

Description automatically generated with low confidence**

This dataset displays the Top 6 Jobs under “Venture Capital & Private Equity” Industryhaving a result of $98K -$188K Salary Estimate to $56K – $102K.

**Now using df\_subset1 I wanted to look at the Bottom 15 Salary Average sorted in ascending order**

As the Salary Estimate column was characters and not numeric I needed to install packages (stingr) so I can extract the digits from the Salary Estimate column.

Then using mutate and str\_ extract from min\_salary and max\_salary in the Salary Estimate I was able to calculate the Salary Average:

**A screenshot of a computer

Description automatically generated with medium confidence**

**I then arranged the Salary Average to show ascending (lowest) number first and print the head of df\_subset1**

**A picture containing text, font, screenshot

Description automatically generated**

**Finally I printed out the bottom 15 of Salary Average**

**A screenshot of a computer program

Description automatically generated with low confidence**

This result is quite interesting and looking at the job title column there are zero Business Analyst jobs in the bottom Salary Average which suggests it pays quite well.

**Now using df\_subset1 I wanted to look at the Top 15 Salary Average sorted in descending order**

A screenshot of a computer screen

Description automatically generated with low confidence

This dataset displays the Top 15 jobs by Salary Average and we can clearly see that Business Analyst Job Title features heavily and accounts for over 80% of the results with $172K Salary Average.

**Now I wanted to create a new df\_subset2 to analyse and focus on three specific columns Job Title, Industry and Salary Average.**

A screenshot of a computer

Description automatically generated with medium confidence

This dataset displays the Top 10 jobs by Salary Average within which Industry and job title. Tech related Industries feature heavily in this list apart from Staff & Outsourcing and Real Estate.

**Now I wanted to group by and count the Industry to see which has the highest by descending order.**

A screenshot of a computer program

Description automatically generated with medium confidence

This dataset displays the Top 10 Industry by count of jobs. Unsurprisingly IT Services is top at 653 and now I wanted to see which Industry was at the bottom.

**df-subset2 group by and count the Industry to see which has the lowest by ascending order.**

A screenshot of a computer code

Description automatically generated with medium confidence

This dataset displays the Bottom 10 Industry by count of jobs. There are no tech related Industries in this list and Commercial Equipment Repair and Maintenance is at the bottom.

**Now add Salary Average column from df-subset1 into the main dataset so I can carry out further analysis.**

I used left join by using the common column “Size” from the df\_subset 1 and the main BusinessAnalyst dataset.

I moved Salary Average into the BusinessAnalyst dataset however it caused duplicate rows of over 2.3 million as the Size column wasn’t a good chose as a common column between the two.

I corrected this using the distinct function on the index column as below:

A picture containing text, font, screenshot

Description automatically generated

I used head BusinessAnalyst to check query to check the Salary Average column had moved across as shown below highlighted in red:

A picture containing text, font, screenshot

Description automatically generated

**Now analysing the Easy Apply column I wanted to count the number of times it is recorded.**

I created a conditional statement to create a new column that displays TRUE when the value of the Easy Apply column is 1 and FALSE when the value of the Easy Apply column is -1.

I then used this new column to group the data and calculate the count for each group.

I used the mutate and ifelse functions in R.

A picture containing text, screenshot, font

Description automatically generated

This dataset displays the Easy\_Apply\_Status count as 134 for True and 3901 as False.

**I wanted to work out the percentage of True across the whole dataset of 4035 so I created the following:**

A picture containing text, font, line, screenshot

Description automatically generated

This dataset displays the Easy\_Apply\_Status percentage of 3.32% across the whole dataset.

**Conclusion**

The report found that:

* Business Analysts were the most common and diverse job title in the dataset, with varying ratings across different industries and sectors.
* Health Care Services and Hospitals was the top industry, and Health Care was the top sector in the dataset, followed by Information Technology and Business Services.
* New York was the most frequent headquarters location in the dataset, followed by Chicago and San Francisco.
* A significant number of companies in the dataset were established between 1980 and 2010, indicating that they are relatively new and may not have a long history or heritage.
* The median rating of companies varied across different sectors, with Information Technology and Education having the highest median ratings above 4 out of 5.
* A significant number of companies in Accounting & Legal, Oil, Gas, Energy and Utilities, and Retail sectors were founded in the years between 1985 and 2000.
* The revenue of companies ranged from less than $1 million to more than $10 billion, with most companies having a revenue between $10 million and $100 million.
* The size of companies ranged from 1 to 10000+ employees, with most companies having between 51 and 200 employees.
* The minimum salary of Business Analysts ranged from $24k to $113k per year, with most salaries between $40k and $80k per year.
* The maximum salary of Business Analysts ranged from $44k to $190k per year, with most salaries between $60k and $100k per year.

The report concluded that:

* The dataset provided a comprehensive overview of the Business Analyst Jobs market in terms of various characteristics and factors that affect the job satisfaction and performance of Business Analysts.
* The dataset revealed some interesting trends and insights about the distribution and correlation of different variables in the dataset, such as industry, sector, rating, revenue, founded year, size, and salary.
* The dataset suggested some areas for further research and analysis, such as exploring the relationship between rating and revenue, or between salary and location.
* The dataset also highlighted some limitations and challenges, such as missing values, outliers, or inconsistencies in some variables that may affect the accuracy and validity of the analysis.

The report recommended that:

* Future data analysis reports should use more advanced techniques and methods to clean, process, and analyse the data more efficiently and effectively.
* Future data analysis reports should also use more interactive and dynamic plot visualizations to enhance the presentation and communication of the findings.
* Future data analysis reports should also include more qualitative data and feedback from Business Analysts themselves to complement the quantitative data and provide a richer and deeper understanding of their experiences and expectations.