A Report on

The data visualization techniques used to explore the features of the dataset

"Data Analyst Salaries"

Submitted by

Ana Das

ID: 2622436

School of Business, University of Dundee

BU51037: Data Visualization for Business

Assessment 02

Dec 6th, 2024

Introduction:

Glassdoor is a leading platform for job seekers and employers, providing insights into work-places, salaries, and career opportunities. It offers company reviews, salary reports, interview questions, and employee benefits information. This data helps job seekers make informed decisions and allows employers to understand market trends and improve their hiring practices. To understand the Business Analytics job market in the USA, a focused segment of Glassdoor data was extracted. This subset includes job listings, estimated high and low salaries, required skills, company locations, and industry sectors etc. The objective was to perform basic analysis to uncover key trends and insights relevant to this field.

Data Exploration:

Originally developed by American mathematician John Tukey in the 1970s, EDA techniques are used to summarizing and understanding the main characteristics of a dataset, often with visual and statistical methods is crucial for uncovering patterns, relationships, anomalies, and potential insights before applying formal modeling. One of the basic ways this approach can be operationalized for examining univariate and bivariate data is by constructing histograms and stem-and-leaf diagrams, box-and-whisker plots, frequency polygons, and cumulative curves (Urban & Wells, 2005).

Research Questions:

- What is the relationship between Industry type and salary, and does this vary by company size?
- What are the salary disparities between major tech hubs (e.g., California, New York) and other regions like Pennsylvania or Texas?

Data Cleaning & Preprocessing:

The initial dataset contained several unnecessary fields, including company name, founding year, company age, industry type, sector, and company description, which were not relevant for the analysis and were therefore removed. Additionally, redundant rows with duplicate data for the same company were eliminated. Extraneous text elements, such as "(Glassdoor est.)" and "(Employer

Est.)," were also removed. These manual preprocessing steps resulted in a clean and streamlined dataset ready for analysis.

Summary of this new dataset:

The dataset (388 rows), gives a brief of different job roles in different states in USA, that is similar to Business Analyst.
Job Title: Roles the company hires for
Seniority: Experience Level
Rating: Rating of the company 1 – 5
Size: Size of the company
Type of Ownership: Is the company Public or Private
Sector: What sector the company belongs to, IT, Fintech etc.,
Lower Salary (K)
Upper Salary (K)
Avg Salary(K)
Job Location: Location of the job

Data Summary & Table Structure:

```
> summary(data)
     index        job_title_sim       Seniority            Rating          Size          Type of ownership     Sector          Lower Salary     Upper Salary
 Min.   :  1.00   Length:388         Length:388         Min.   :1.900   Length:388         Length:388         Length:388         Min.   : 15.00   Min.   : 16.00
 1st Qu.: 98.75   Class :character   Class :character   1st Qu.:3.400   Class :character   Class :character   Class :character   1st Qu.: 56.00   1st Qu.: 99.75
 Median :195.50   Mode  :character   Mode  :character   Median :3.800   Mode  :character   Mode  :character   Mode  :character   Median : 72.00   Median :126.00
 Mean   :195.40                                         Mean   :3.769                                                            Mean   : 76.41   Mean   :130.11
 3rd Qu.:292.25                                         3rd Qu.:4.125                                                            3rd Qu.: 92.00   3rd Qu.:153.00
 Max.   :389.00                                         Max.   :5.000                                                            Max.   :202.00   Max.   :306.00
 Avg Salary(K)    Job Location          Python            spark             aws             excel             sql             hadoop            tableau
 Min.   : 15.5   Length:388         Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 77.5   Class :character   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :100.0   Mode  :character   Median :1.0000   Median :0.0000   Median :0.0000   Median :1.0000   Median :1.0000   Median :0.0000   Median :0.0000
 Mean   :103.3                      Mean   :0.6392   Mean   :0.2835   Mean   :0.2577   Mean   :0.5464   Mean   :0.6418   Mean   :0.2088   Mean   :0.2448
 3rd Qu.:123.5                      3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :254.0                      Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
>
```

```
> str(data)
tibble [388 × 18] (S3: tbl_df/tbl/data.frame)
 $ index            : num [1:388] 70 267 170 337 165 273 150 92 328 202 ...
 $ job_title_sim    : chr [1:388] "Business Analyst" "Business Analyst" "Business Analyst" "Business Analyst" ...
 $ Seniority        : chr [1:388] "Junior" "Associate" "Associate" "Associate" ...
 $ Rating           : num [1:388] 4.7 4.6 4.3 3.8 3.8 3.7 4.7 4.3 3.3 4.2 ...
 $ Size             : chr [1:388] "201 - 500" "501 - 1000" "1001 - 5000" "1001 - 5000" ...
 $ Type of ownership: chr [1:388] "Private" "Private" "Private" "Private" ...
 $ Sector           : chr [1:388] "Advertising & Marketing" "Finance" "Transportation & Logistics" "Travel & Tourism" ...
 $ Lower Salary     : num [1:388] 37 31 53 36 44 49 42 39 90 37 ...
 $ Upper Salary     : num [1:388] 76 55 105 71 86 76 76 71 157 68 ...
 $ Avg Salary(K)    : num [1:388] 56.5 43 79 53.5 65 ...
 $ Job Location     : chr [1:388] "PA" "IA" "OH" "NY" ...
 $ Python           : num [1:388] 0 0 0 1 0 0 1 0 0 1 ...
 $ spark            : num [1:388] 0 0 0 0 0 0 0 0 0 0 ...
 $ aws              : num [1:388] 0 0 0 0 0 0 0 0 0 0 ...
 $ excel            : num [1:388] 1 1 0 1 1 1 1 1 1 0 ...
 $ sql              : num [1:388] 0 1 1 1 1 0 0 1 1 1 ...
 $ hadoop           : num [1:388] 0 0 0 0 0 0 0 0 0 0 ...
 $ tableau          : num [1:388] 0 1 0 0 1 0 0 0 1 0 ...
```

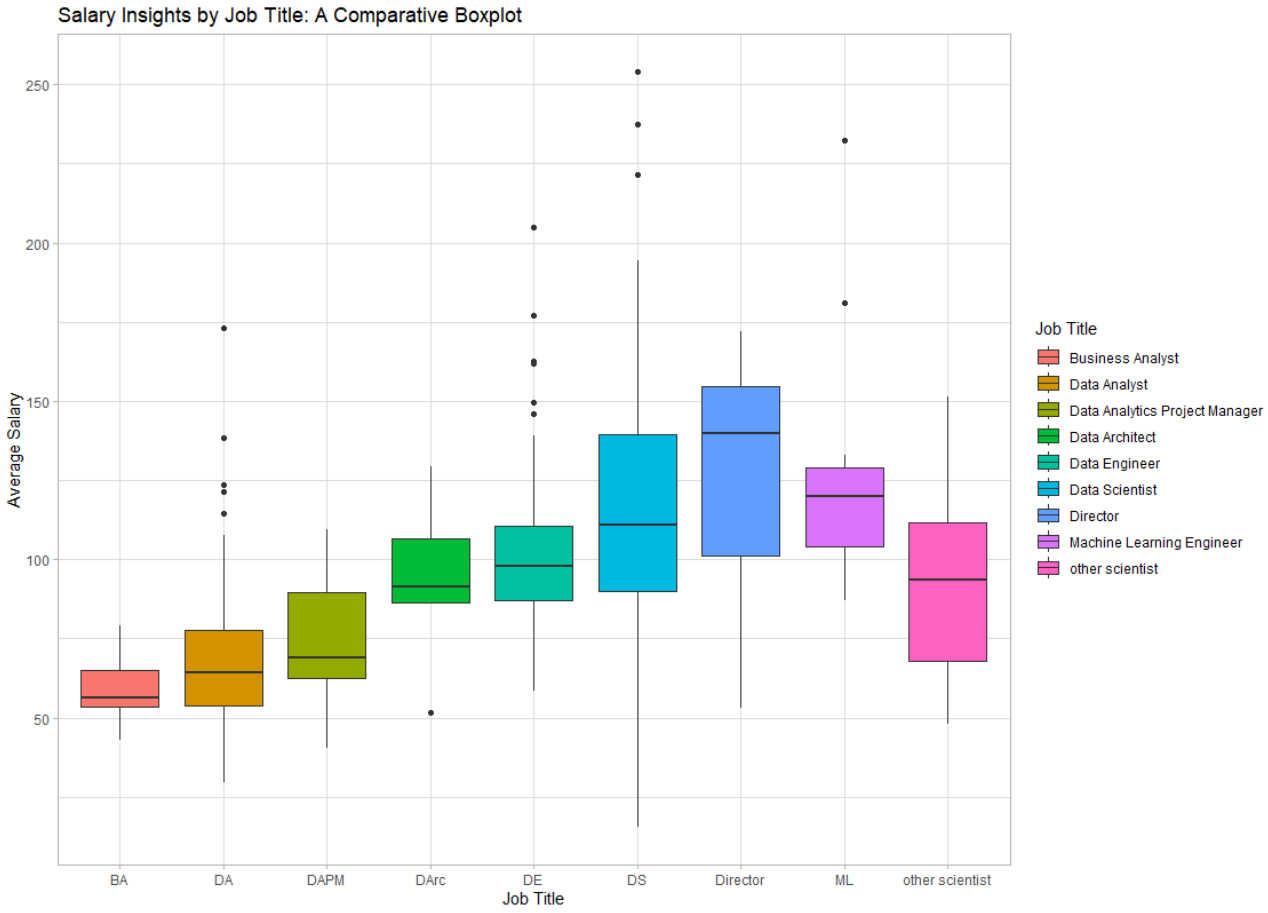
Check for Missing Values:

```
> any(is.na(data))
[1] FALSE
```

Data Visualization:

1. Box Plot:

A box plot, also known as a box-and-whisker plot, offers a concise visualization of the distribution, central tendency, and variability within a dataset, making it particularly effective for comparing distributions across categories (McGill et al., 1978). In this analysis, the box plot illustrates average salary distributions by job title, highlighting salary ranges, medians, and outliers for each role. This visualization is particularly useful for identifying outliers and facilitating group comparisons. By revealing the salary range for various job titles, the plot aids job seekers in setting realistic salary expectations and identifying roles with higher earning potential. Furthermore, the box plot provides organizations with actionable insights into salary structures across job roles, enabling benchmarking of compensation packages, cost optimization, and talent acquisition strategies. It sheds light on market trends, the demand for specific positions, and disparities in pay, thereby supporting strategic workforce planning and the promotion of equitable and competitive compensation practices.

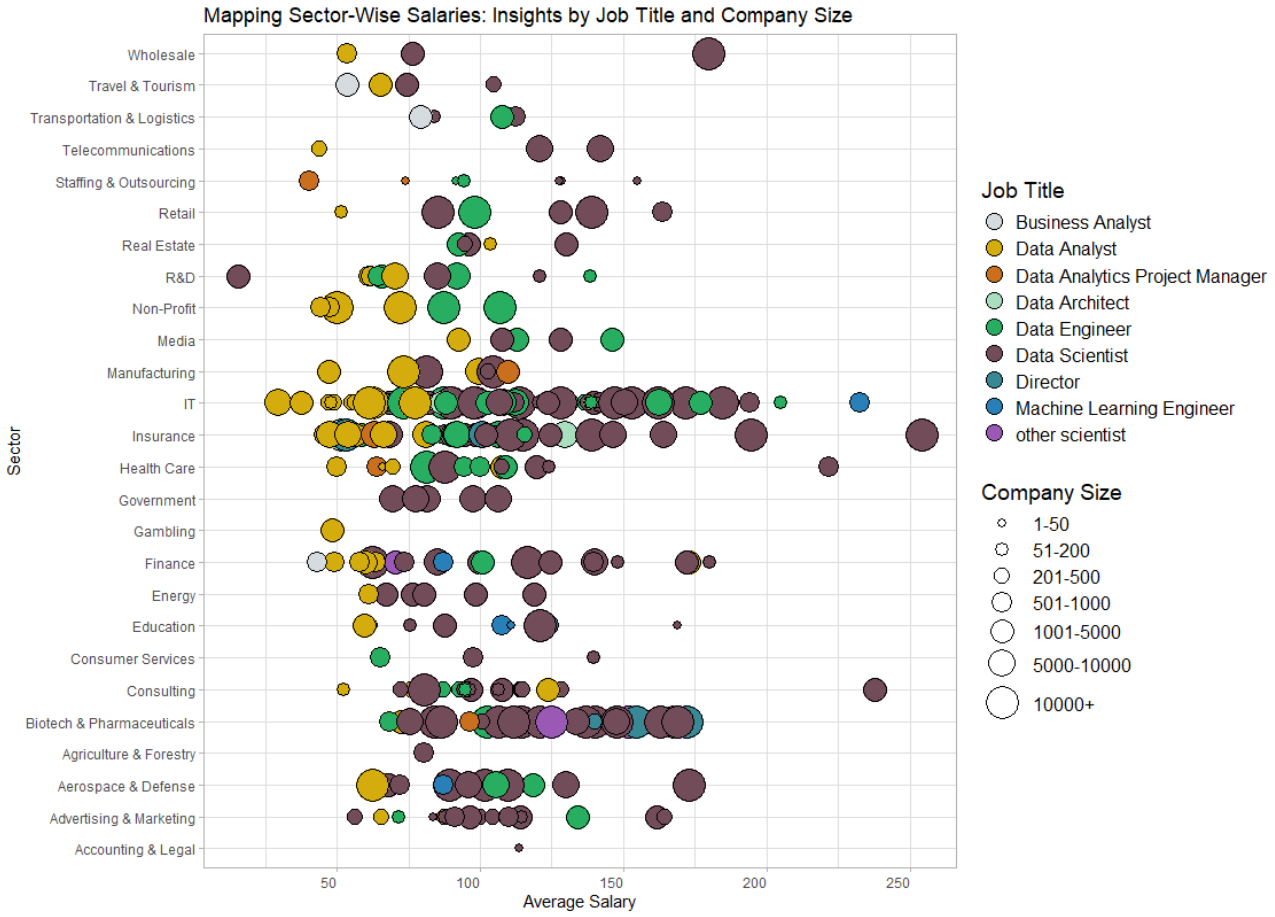


Outlier Detection and Impact:

While extreme values in Avg_Salary, Lower_Salary, and Upper_Salary have been observed, these outliers reflect legitimate high-paying jobs rather than data errors. However, these extreme salary values can skew the overall salary average, potentially distorting trends in the dataset (Wang, 2021). For instance, a small number of very high salaries can inflate the mean, making it appear higher than the majority of jobs. Due to limitations in the dataset, such as the absence of experience data and a relatively small sample size of 388 entries, it is difficult to accurately assess the impact of these outliers.

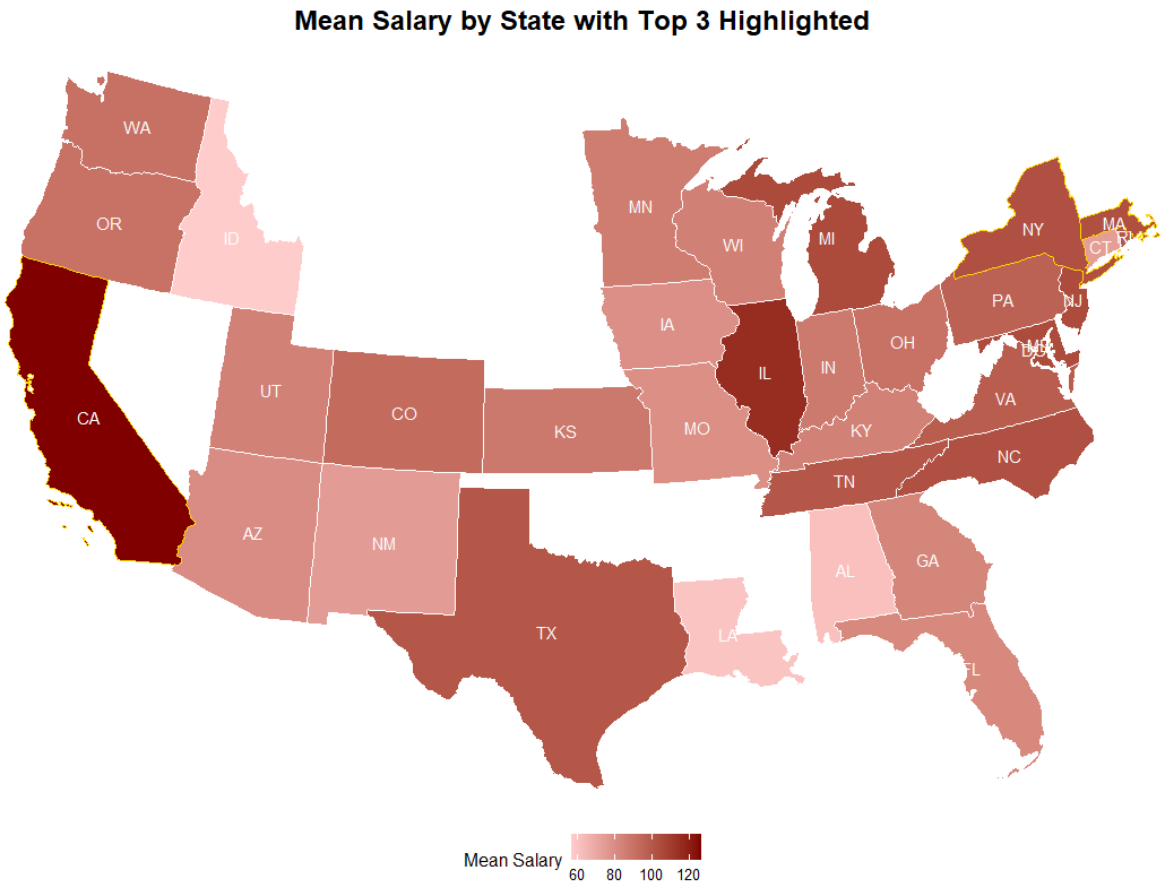
2. Bubble plot:

A bubble plot is an advanced data visualization tool used to represent three dimensions of data on a two-dimensional plane. It extends the standard scatter plot by incorporating bubble size as an additional variable (Bessler, 2023). In this context, the plot is utilized to illustrate the relationships among four dis-tinct variables (Average Salary, Sector, Company Size, and Job Title) simultaneously, highlighting how salary is influenced by these market factors. Comparisons of sectors, job titles, and company sizes with respect to salaries are visually facilitated by the plot. The application of a logarithmic scale on the X-axis ensures effective management of the wide range of salary data. Employers are provided with insights to benchmark their compensation packages against competitors within the same sector or company size category. Similarly, sectors and job titles offering higher average salaries can be identified by job seekers, aiding in career decision-making and aligning opportunities with their skills.



3. US Choropleth Map:

A Choropleth map is a thematic map where regions are shaded based on a variable's value, making it ideal for illustrating the spatial distribution of data, revealing patterns or disparities across geographic areas (Chien et al., 2019). In the context of the dataset containing state-wise salary data, the map highlights regions with higher or lower average salaries, helping job seekers identify lucrative opportunities and businesses pinpoint areas with varying salary expectations. This allows job seekers to focus their search on high-paying states, while businesses can adjust recruitment strategies and compensation packages. Additionally, companies can assess regional salary disparities to attract top talent by expanding into competitive wage areas.



Analysis & Comprehensive Data Summary:

The salary analysis across roles, locations, sectors, and skills reveals significant patterns in the U.S. job market, highlighting the evolving demands of various industries and the value of specific competencies.

Salaries escalate with seniority, ranging from $54K–$92K for junior roles to $94K–$156K for senior-level positions. Among specific roles, Machine Learning Engineers command the highest average salary of $128K, but their market demand is relatively limited. In contrast, Data Scientists, with an average salary of $116K, are in high demand and represent one of the most sought-after roles. Data Engineers, averaging $102K, are increasingly indispensable due to the rising implementation of AI models, particularly in California. Meanwhile, Data Analysts, earning an average of $71K, are experiencing growing demand in key locations like New York and California. Business Analysts, on the other hand, are the lowest-paid, averaging $59K, with most opportunities at the associate level. (Box Plot)

The Biotech sector offers the highest average salary at $124K, underscoring its lucrative and specialized nature. IT and Advertising closely follow at $108K, driven by digital transformation and creative innovations. Sectors like Consulting, Insurance, Finance, and Healthcare maintain competetive averages of $100K–$104K, while Manufacturing ($89K) and Energy ($81K) lag behind, reflecting traditional industry challenges. (Bubble Plot)

Location plays a critical role in salary variations. California leads with an average salary of $126K, reflecting its status as a tech and innovation hub. Illinois follows at $115K, while Massachusetts, New York, and Maryland average around $102K. States like Virginia ($98K), Texas ($100K), and Washington ($90K) show modest averages, highlighting regional disparities driven by local industries and cost of living. (Choropleth Map)

Technical expertise remains pivotal, with professionals proficient in Python, SQL, and Excel earning an average of $102K. These skills align closely with roles in data science and engineering, underscoring their critical importance across sectors.

Limitations of the Current Dataset:

The dataset includes categorical variables, such as Job_Title, which have a broad range and may require further breakdown for more meaningful analysis. The Seniority data appears vague, making it challenging to identify trends based on experience. Additionally, the dataset focuses primarily on salaries in specific locations (states), limiting its ability to reflect global or international trends. Certain industries are underrepresented, potentially skewing the results. Moreover, the dataset may not capture all forms of compensation, such as bonuses, stock options, or other non-salary benefits, which could provide a more comprehensive view of overall compensation. These areas could be considered for further exploration in future studies.

Areas for Further Exploration:

Future investigations could explore salary trends across different job titles and seniority levels, with a deeper focus on how salaries vary based on job title and experience. Additionally, the relationship between skills such as Python, Spark, AWS, and SQL with salary could be examined to

determine whether certain skill sets enhance salary prospects. An analysis of how salary trends in the data industry have evolved over the past five years, along with the factors driving these changes, would provide valuable insights. Finally, gender-based salary disparities should be considered to assess potential inequalities in compensation.

Conclusion:

This comprehensive analysis illustrates the interconnected dynamics of roles, locations, sectors, and skills, revealing the factors driving salaries in the U.S. job market. Specialized roles and industries, technical skills, and geographical advantages are key contributors to higher earnings, painting a picture of a job market shaped by innovation, demand, and regional focus (Septiandri et al., 2024). By analysing relationships between the variables, meaningful trends can be identified that benefit both job seekers and businesses. However, attention must be given to data quality, missing values, outliers, and potential biases to ensure the results are valid and actionable. Identifying the limitations and areas for further exploration will guide deeper analysis, making the findings more robust and applicable to real-world decision-making.

Appendix:

Some Interesting Facts about the Dataset:

Small companies (1–50 employees) offer the highest average salary at $112K, but tend to require fewer specialized skills, focusing on Python, Excel, and SQL. These firms, mostly private, tend to hire more associates than seniors, and predominantly located in California (CA) and Pennsylvania (PA). Their salary correlates with company rating, with those rated higher paying more. As companies grow in size, salaries decrease slightly, with medium-sized firms (50–200 employees) offering $108K on average. These companies also focus on associate-level roles, with a similar skill set but a slightly higher rating of 4. Large enterprises (1001–5000 employees) see a further drop in average salary to $99K, and while they still prioritize Python and SQL, they also demand broader technical skills like Spark, AWS, and Tableau. In very large enterprises (5001–10,000 employees) and corporate giants (10,000+ employees), salaries rise again, with the latter offering the highest average salary at $115K. These companies also place a stronger emphasis on senior-level hires. Interestingly,

within public companies, higher ratings do not always translate to higher pay, as seen in larger firms where a rating of 3–3.9 can yield higher salaries than those rated 4–4.9. Surprisingly, Health, Beauty, & Fitness shows unexpectedly high salaries ($139.5K). Tech hubs (CA) command premium salaries.

Codes:

```
library (ggplot2)
library (readxl)
library (tidyverse)
library(dplyr)
library(ggally)
library(ggcorrplot)
#install.packages("ggcorrplot")
library(usmap)
library(sf)
library(ggalt)
library(ggrepel)

view(data)
summary(data)
head(data)
str(data)
sapply(data, class)
any(is.na(data))
```

Box Plot:

```
data <- data %>% arrange(Sector, Ownership) %>% group_by(Job_Title) %>% mutate(Avg_Salary) %>% ungroup()
ggplot(data, aes(x = Job_Title, y = Avg_Salary, fill = Job_Title)) +
  geom_boxplot() +
  labs(
    title = "Salary Insights by Job Title: A Comparative Boxplot",
    x = "Job Title",
    y = "Average Salary",
    fill = "Job Title"
  ) +
  theme_light() +
  scale_x_discrete(labels = c(
    "Data Scientist" = "DS",
    "Business Analyst" = "BA",
    "Data Engineer" = "DE",
    "Marketing Manager" = "MM",
    "Data Analyst" = "DA",
    "Data Analytics Project Manager" = "DAPM",
    "Data Architect" = "DArc",
    "Machine Learning Engineer" = "ML"
  ))
```

Bubble Plot:

```
data$Company_Size_Group <- cut(data$Company_Size,
                               breaks = c(0, 50, 200, 500, 1000, 5000, 10000, Inf),
                               labels = c("1-50", "51-200", "201-500", "501-1000", "1001-5000", "5000-10000", "10000+"))

ggplot(data, aes(x = Avg_Salary, y = Sector, size = Company_Size_Group, fill = Job_Title)) +
  geom_point(shape = 21, color = "black") +
  # scale_x_log10() +
  scale_size_manual(values = c(2.5, 3.8, 4.5, 6, 7, 8, 10)) +
  scale_fill_manual(values = c(
    "#e6b0aa", "#d98880", "#cd6155", "#229954", "#2ecc71", "#5499c7", "#40b0c1", "#2e86c1",     "#801a1a", "#dc7633" ))
+
  labs(title = "Mapping Sector-Wise Salaries: Insights by Job Title and Company Size",
       x = "Average Salary",
       y = "Sector",
       size = "Company Size",
       fill = "Job Title") +
```

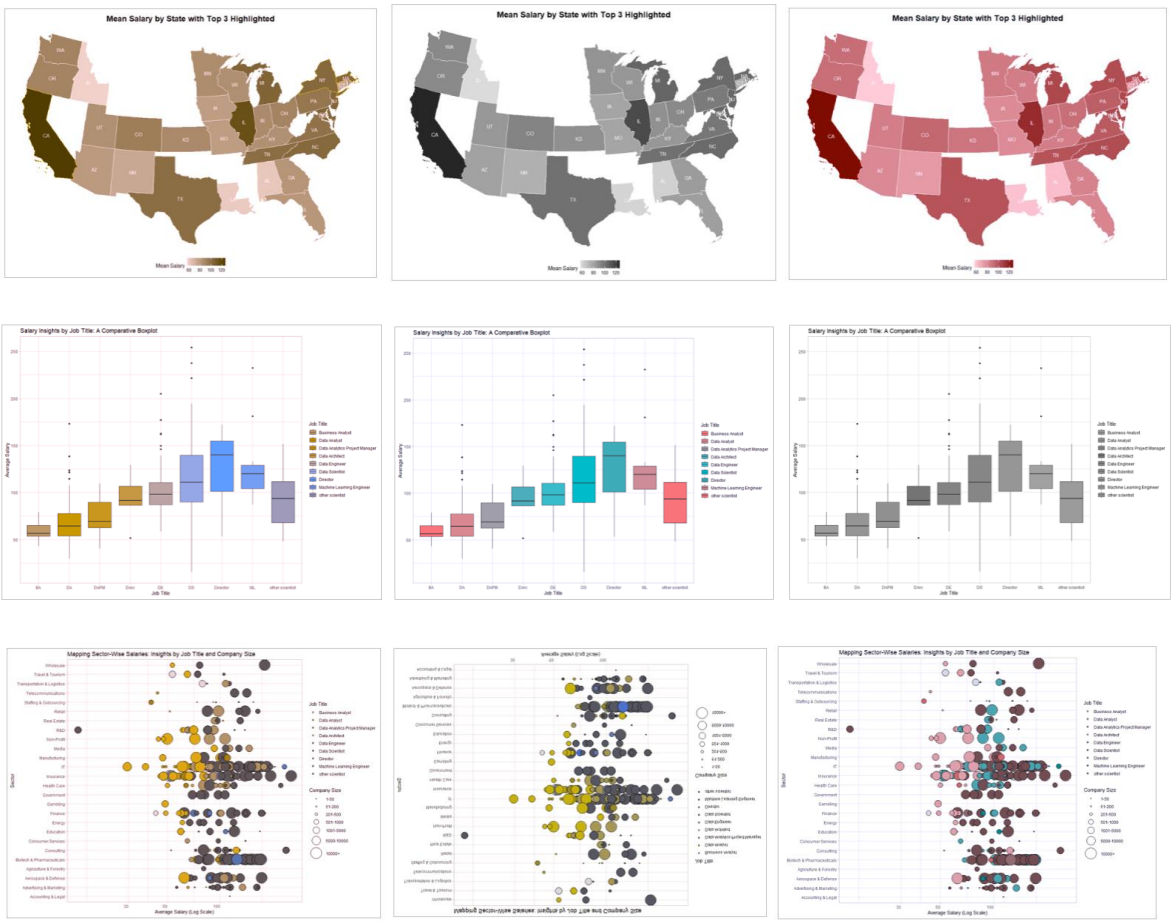
```
  theme_light() +
  theme(axis.text.x = element_text(hjust = 1) ,
        legend.text = element_text(size = 12),  # Adjust size of legend text
        legend.title = element_text(size = 14)) +  # Adjust size of legend title
  guides(fill = guide_legend(
    override.aes = list(size = 5) # Set the size of the legend bubbles for Job Title
  ))
```

Choropleth Map:

```
state_salaryx <- data %>%
  group_by(Job_Location) %>%
  summarise(
    Mean_Salary = mean(Avg_Salary, na.rm = TRUE),
    Salary_Count = sum(!is.na(Avg_Salary))
  )
map_data <- us_map(regions = "states")

map_data <- map_data %>%
  left_join(state_salaryx, by = c("abbr" = "Job_Location"))

summary(map_data$Mean_Salary)
map_data <- map_data %>%
  filter(!is.na(Mean_Salary))

view(map_data)

  top_3_states <- map_data %>%
  arrange(desc(Salary_Count)) %>%
  slice_head(n = 3) %>%
  pull(abbr)  # Get the top 3 state abbreviations

map_data <- map_data %>%
  mutate(highlight = ifelse(abbr %in% top_3_states, "Highlight", "Normal"))

state_centroids <- st_centroid(map_data)

ggplot(data = map_data) +
  geom_sf(aes(fill = Mean_Salary), color = "white") +
  geom_sf(data = map_data %>% filter(highlight == "Highlight"),
          aes(fill = Mean_Salary), color = "gold", size = 40, alpha = 40) +
  scale_fill_gradient(low = "#ffcccc", high = "#800000", na.value = "grey90") +
  geom_sf_text(data = state_centroids, aes(label = abbr), size = 4, color = "white") +
  labs(title = "Average Salary by State with Top 3 States Highlighted", fill = "Mean Salary") +
  theme_void() +
  theme(
    plot.title = element_text(size = 18, face = "bold", hjust = 0.5),
    legend.position = "bottom",
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10)
  )
```

Colour Sources: https://htmlcolorcodes.com/

All the graphs are visually effective for colorblind people: Color Blind Simulator



References:

Tukey, J. W. (1970). *Exploratory Data Analysis. United States: Addison Wesley Publishing Company*.

Urban, P. A., & Wells, C. (2005). *Archaeology. Encyclopedia of Social Measurement*, 71–81. https://www.sciencedirect.com/science/article/pii/B0123693985003169

Mcgill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, *32*(1), 12–16. https://doi.org/10.1080/00031305.1978.10479236

Wang, D. (2021). *The Impact of Outliers on Regression Coefficients: A Sensitivity Analysis*. The International Journal of Accounting, 56(3). https://doi.org/10.1142/S1094406021500141

Bessler, L. (2023). *Bubble Plots. In Visual Data Insights Using SAS ODS Graphics* (pp. 263–281). Apress. https://doi.org/10.1007/978-1-4842-8609-8_7

Chien, T.-W., Wang, H.-Y., Hsu, C.-F., & Kuo, S.-C. (2019). *Choropleth map legend design for visualizing the most influential areas in article citation disparities: A bibliometric study*. *Medicine (Baltimore)*, *98*(41), e17527–e17527. https://doi.org/10.1097/MD.0000000000017527

Septiandri, A. A., Constantinides, M., & Quercia, D. (2024). *The potential impact of AI innovations on US occupations*. *PNAS Nexus*, *3*(9). https://doi.org/10.1093/pnasnexus/pgae320