BUSINESS ANALYST CAREERS DATASET

Given: BusinessAnalyst.xlsx

Fields: Unnamed: 0, index, Job Title, Salary Estimate, Job Description, Rating, Company Name, Location, Headquarters, Size, Founded, Type of ownership, Industry, Sector, Revenue, Competitors, Easy Apply.

	Unnamed: 0	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	
0	0	0	Business Analyst - Clinical & Logistics Platform	56K–102K (Glassdoor est.)	Company Overview\n\n\nAt Memorial Sloan Ketter...	3.9	Memorial Sloan-Kettering\n3.9	New York, NY	New York, NY	10000+ employees	1884	Nonprofit Organization	Health Care Services & Hospitals	Healt
1	1	1	Business Analyst	56K–102K (Glassdoor est.)	We are seeking for an energetic and collaborat...	3.8	Paine Schwartz Partners\n3.8	New York, NY	New York, NY	1 to 50 employees	-1	Company - Private	Venture Capital & Private Equity	F
2	2	2	Data Analyst	56K–102K (Glassdoor est.)	For more than a decade, Asembia has been worki...	3.6	Asembia\n3.6	Florham Park, NJ	Florham Park, NJ	501 to 1000 employees	2004	Company - Private	Biotech & Pharmaceuticals	Bio Pharmace
3	3	3	Information Security Analyst, Incident Response	56K–102K (Glassdoor est.)	Job Description Summary\nThe Information Secur...	3.6	BD\n3.6	Franklin Lakes, NJ	Franklin Lakes, NJ	10000+ employees	1897	Company - Public	Health Care Products Manufacturing	Manufa
4	4	4	Analyst - FP&A Global Revenue	56K–102K (Glassdoor est.)	Magnite is the world's largest independent sel...	3.4	Rubicon Project\n3.4	New York, NY	Los Angeles, CA	201 to 500 employees	2007	Company - Public	Internet	Infor Techi

Fig. 1. First six records of the dataset

Dataset Cleaning:

I used Python to read the given dataset as a data frame into Jupyter Notebook for cleaning. There are 4092 records (rows) and 17 columns. In addition to two columns not needed (Unnamed: 0 and index), the dataset is misaligned from a particular index number to the end of the dataset. By means of two functions: is_aligned() and fix_misaligned(), this is corrected by shifting the dataset to the right.

```
import pandas as pd

def is_misaligned(row):
    # Check if 'Unnamed: 0' is not a number (it should be an index)
    unnamed_0 = row['Unnamed: 0']
    index_val = row['index']

    if isinstance(unnamed_0, str) and not unnamed_0.isdigit():
        # Check if 'index' contains a salary pattern like '102K'
        if isinstance(index_val, str) and 'K' in index_val:
            return True
    return False
```

```
def fix_misaligned_row(row):
    if is_misaligned(row):
        shifted = [pd.NA] * 2 + row.tolist()[:-2]  # shift right by 2
        return pd.Series(shifted, index=row.index)
    return row

df = df.apply(fix_misaligned_row, axis=1)
df
```

Fig 2. Two functions to fix the misalignments in the dataset.

The sample records after fixing the misalignment.

```
[16]: df2 = df2.apply(fix_misaligned_row, axis=1)
      df2.head()
```

[16]:

	Unnamed: 0	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	
0	0	0	Business Analyst - Clinical & Logistics Platform	56K— 102K (Glassdoor est.)	Company Overview\n\n\nAt Memorial Sloan Ketter...	3.9	Memorial Sloan-Kettering\n3.9	New York, NY	New York, NY	10000+ employees	1884	Nonprofit Organization	Health Care Services & Hospitals	Healt
1	1	1	Business Analyst	56K— 102K (Glassdoor est.)	We are seeking for an energetic and collaborat...	3.8	Paine Schwartz Partners\n3.8	New York, NY	New York, NY	1 to 50 employees	-1	Company - Private	Venture Capital & Private Equity	F
2	2	2	Data Analyst	56K— 102K (Glassdoor est.)	For more than a decade, Asembia has been worki...	3.6	Asembia\n3.6	Florham Park, NJ	Florham Park, NJ	501 to 1000 employees	2004	Company - Private	Biotech & Pharmaceuticals	Bio Pharmace
3	3	3	Information Security Analyst, Incident Response	56K— 102K (Glassdoor est.)	Job Description Summary\nThe Information Secur...	3.6	BD\n3.6	Franklin Lakes, NJ	Franklin Lakes, NJ	10000+ employees	1897	Company - Public	Health Care Products Manufacturing	Manufa
4	4	4	Analyst - FP&A Global Revenue	56K— 102K (Glassdoor est.)	Magnite is the world's largest independent sel...	3.4	Rubicon Project\n3.4	New York, NY	Los Angeles, CA	201 to 500 employees	2007	Company - Public	Internet	Infor Techi

Fig. 3. First six records after fixing this misalignment.

Then Unnamed:0 and index fields were dropped. Examining Rating closely revealed that there are 356 records have the value of -1. These records were removed because examining them showed that they had -1 too in other fields. Now our records remained 3736 with the index reset. We did not lose so much, and besides, these records were meaningless.

Duplicates were removed, no duplicates found. Next, the fields: Location, Headquarters, Type of ownership, Founded, Sector, Revenue, Competitors, and Size had -1, 'Unknown', 'Unknown / Non-Applicable' as values. These were replaced with NAN. But the field 'Easy Apply' had 1 and -1. I converted the -1 to 0, so that the field has 1 and 0 for Yes and No respectively. The field was the converted to int.

I added another field called Job_Category to categorize the Job Title into categories: Data Analyst, BI Analyst, Business Analyst, Marketing Analyst, Finance Analyst, Technical Analyst, Product Analyst, Operations Analyst, Research Analyst, ERP Analyst, other Analyst and Non-Analyst. Two more fields were added: salary_lower and salary_upper. These fields were based on the Salary_Estimate and would help in analysis. The following images show the dataset information and summary statistics of the numeric fields.

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3736 entries, 0 to 3735
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Job Title          3736 non-null   object 
 1   Salary Estimate    3736 non-null   object 
 2   Job Description    3736 non-null   object 
 3   Rating             3736 non-null   float64
 4   Company Name       3736 non-null   object 
 5   Location           3736 non-null   object 
 6   Headquarters       3726 non-null   object 
 7   Size               3690 non-null   object 
 8   Founded            3007 non-null   Int64  
 9   Type of ownership  3704 non-null   object 
 10  Industry           3531 non-null   object 
 11  Sector             3531 non-null   object 
 12  Revenue            2909 non-null   object 
 13  Competitors        1105 non-null   object 
 14  Job_Category       3736 non-null   object 
 15  salary_lower       3736 non-null   int64  
 16  salary_upper       3736 non-null   int64  
 17  Easy Apply         3736 non-null   int32  
dtypes: Int64(1), float64(1), int32(1), int64(2), object(13)
memory usage: 514.6+ KB
```

Fig. 4. The information about the dataset.

```
[62]: df2.describe()
```

[62]:

	Rating	Founded	salary_lower	salary_upper	Easy Apply
count	3736.000000	3007.0	3736.000000	3736.000000	3736.000000
mean	3.760278	1975.044563	55122.323340	97911.937901	0.035064
std	0.652571	48.86409	20165.029295	32364.796630	0.183967
min	1.000000	1690.0	27000.000000	48000.000000	0.000000
25%	3.400000	1966.0	41000.000000	78000.000000	0.000000
50%	3.700000	1995.0	48000.000000	87000.000000	0.000000
75%	4.100000	2004.0	63000.000000	111000.000000	0.000000
max	5.000000	2020.0	124000.000000	226000.000000	1.000000

Fig. 5. The summary statistics of the numeric fields.

The cleaned dataset was saved as ‘updated_cleaned_business_analyst.xlsx’.

Analysis in R Studio:

Load the dataset

df <- read_excel('cleaned_business_analyst.xlsx')

This loads the dataset using the variable df

View the structure and data types

glimpse(df)

shows: Rows: 3,736, Columns: 18.

Summary(df)

Shows

```
> summary(df)
  Job Title          Salary Estimate    Job Description        Rating       Company Name         Location          Headquarters
 Length:3736        Length:3736        Length:3736        Min.   :1.00   Length:3736        Length:3736        Length:3736
 Class :character   Class :character   Class :character   1st Qu.:3.40   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :3.70   Mode  :character   Mode  :character   Mode  :character
                                                          Mean   :3.76
                                                          3rd Qu.:4.10
                                                          Max.   :5.00

     Size              Founded     Type of ownership    Industry            Sector             Revenue           Competitors
 Length:3736        Min.   :1690   Length:3736        Length:3736        Length:3736        Length:3736        Length:3736
 Class :character   1st Qu.:1966   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Median :1995   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
                    Mean   :1975
                    3rd Qu.:2004
                    Max.   :2020
                    NA's   :729
 Job_Category        salary_lower     salary_upper      Easy Apply
 Length:3736        Min.   : 27000   Min.   : 48000   Min.   :0.00000
 Class :character   1st Qu.: 41000   1st Qu.: 78000   1st Qu.:0.00000
 Mode  :character   Median : 48000   Median : 87000   Median :0.00000
                    Mean   : 55122   Mean   : 97912   Mean   :0.03506
                    3rd Qu.: 63000   3rd Qu.:111000   3rd Qu.:0.00000
                    Max.   :124000   Max.   :226000   Max.   :1.00000
```

Fig. 6. Summary statistics in R Studio

Fig. 6 gives information about the dataset and the summary statistics for the numeric fields. Rating, for example, shows that the minimum is 1 and maximum is 5, mean is 3.76, first quartile is 3.40, median is 3.70, and third quartile is 4.10.

colSums(is.na(df))

```
> colSums(is.na(df))
        Job Title   Salary Estimate   Job Description            Rating      Company Name          Location      Headquarters              Size
                0                 0                 0                 0                 0                 0                10                46
          Founded Type of ownership          Industry            Sector           Revenue       Competitors      Job_Category      salary_lower
              729                32               205               205               827              2631                 0                 0
     salary_upper        Easy Apply
                0                 0
>
```

Fig. 7. Showing the missing values.

Fig. 7 shows the fields with missing values. Headquarters, 10 values missing, Size, 46 values, Founded, 729, and so forth.

Exploratory Data Analysis

One Variable

```
# Count the number of Job Categories
 > df %>%
+   count(Job_Category) %>%                          # Count occurrences of each title
+   filter(!is.na(`Job_Category`)) %>%              # Remove missing values
+   top_n(5, n) %>%                                  # Show top 15 job titles by count
+   ggplot(aes(x = fct_reorder(`Job_Category`, n), y = n)) +
+   geom_col(fill = "steelblue") +                  # Bar plot
+   coord_flip() +                                   # Flip axes for readability
+   labs(
+     title = "Top 5 Job Titles",
+     x = "Job Title",
+     y = "Number of Jobs"
+   )
```

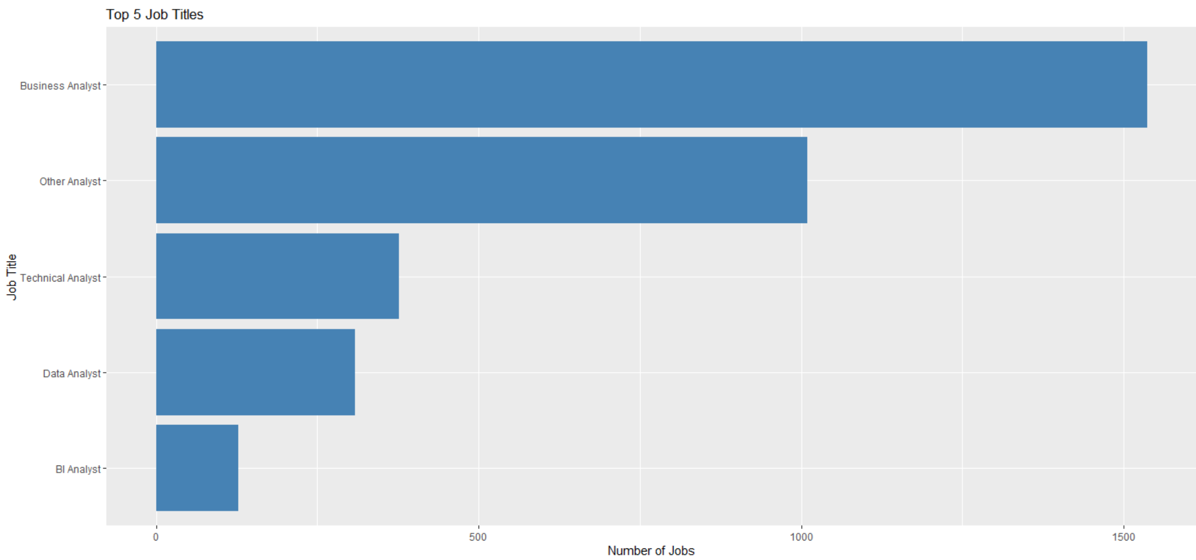

Fig. 8. Top 5 Job Categories

On top is the Business Analyst, followed by Other Analyst, Technical Analyst, Data Analyst, and BI Analyst.

```
df %>%
 count(Industry) %>%              # Count jobs per industry
 filter(!is.na(Industry)) %>%      # Remove missing values
 top_n(5, n) %>%                  # Select top 5 industries by count
 ggplot(aes(x = fct_reorder(Industry, n), y = n)) +
 geom_col(fill = "steelblue") +
 coord_flip() +
 labs(
  title = "Top 5 Industries by Job Count",
  x = "Industry",
  y = "Number of Jobs"
 )
```

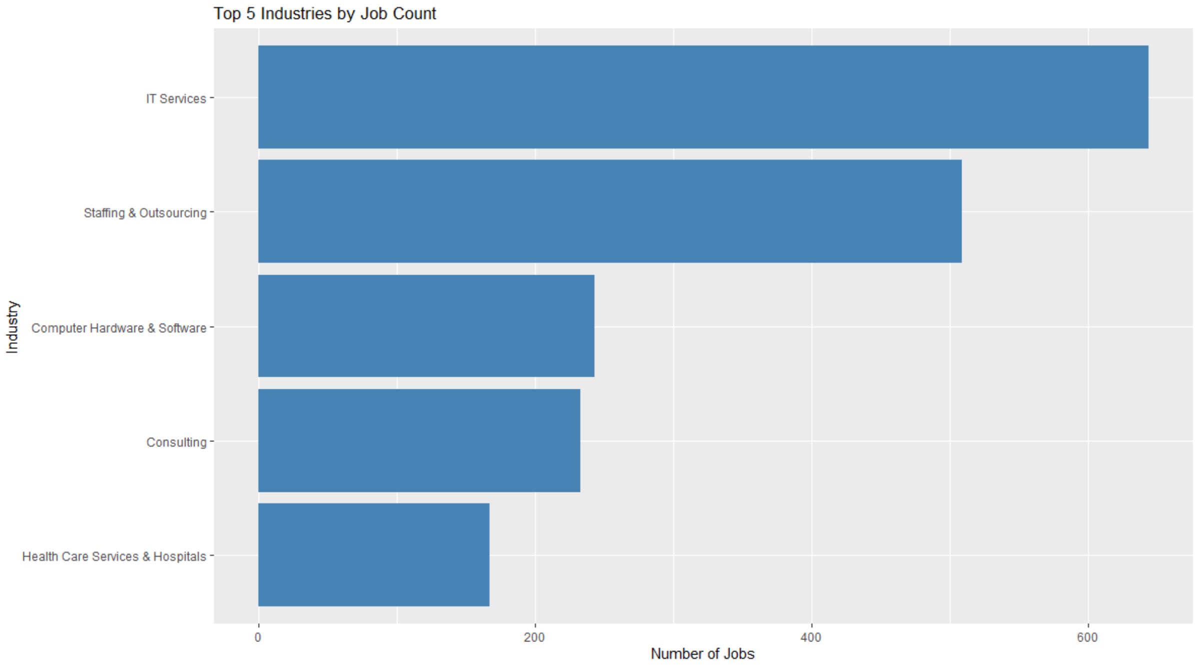

Fig. 9. Top 5 industries in the Data Analyst dataset.

Fig. 9 shows top 5 industries. They are IT Services, Staffing & Outsourcing, Computer Hardware & Software, Consulting, and Health Care Services & Hospitals.

Top 5 locations for Data Analyst Jobs

```
df %>%
 count(Location) %>%
 filter(!is.na(Location)) %>%
 top_n(5, n) %>%
 ggplot(aes(x=fct_reorder(Location, n), y=n)) +
 geom_col(fill="steelblue") + coord_flip() +
 labs(
  title = "Top 5 Job Locations",
  x = "Locations",
  y = "Counts"
 ) +
 theme_minimal()
```

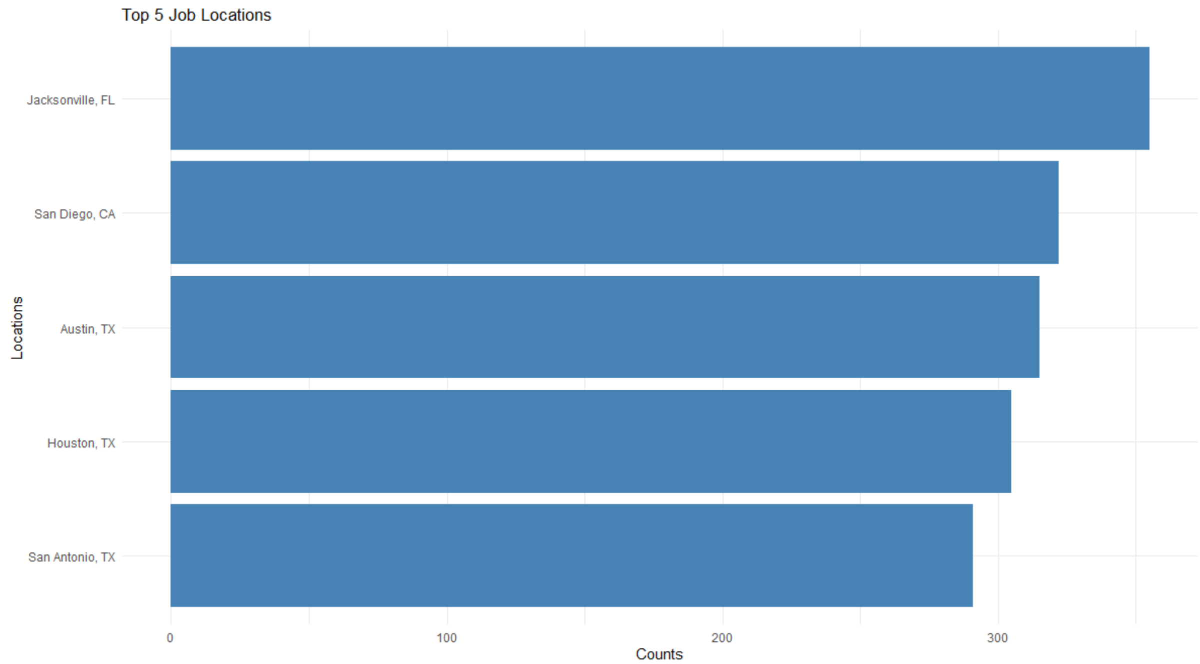

Fig.10. Top 5 Job Locations

The image above shows the top 5 cities for Data Analyst Jobs. Jacksonville, FL, tops, followed by San Diego, CA. In the fifth place is San Antonio, Tx.

Histogram of Rating

```
ggplot(df, aes(x = Rating)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  labs(title = "Histogram of Company Ratings", x = "Rating", y = "Frequency") +
  theme_minimal()
```

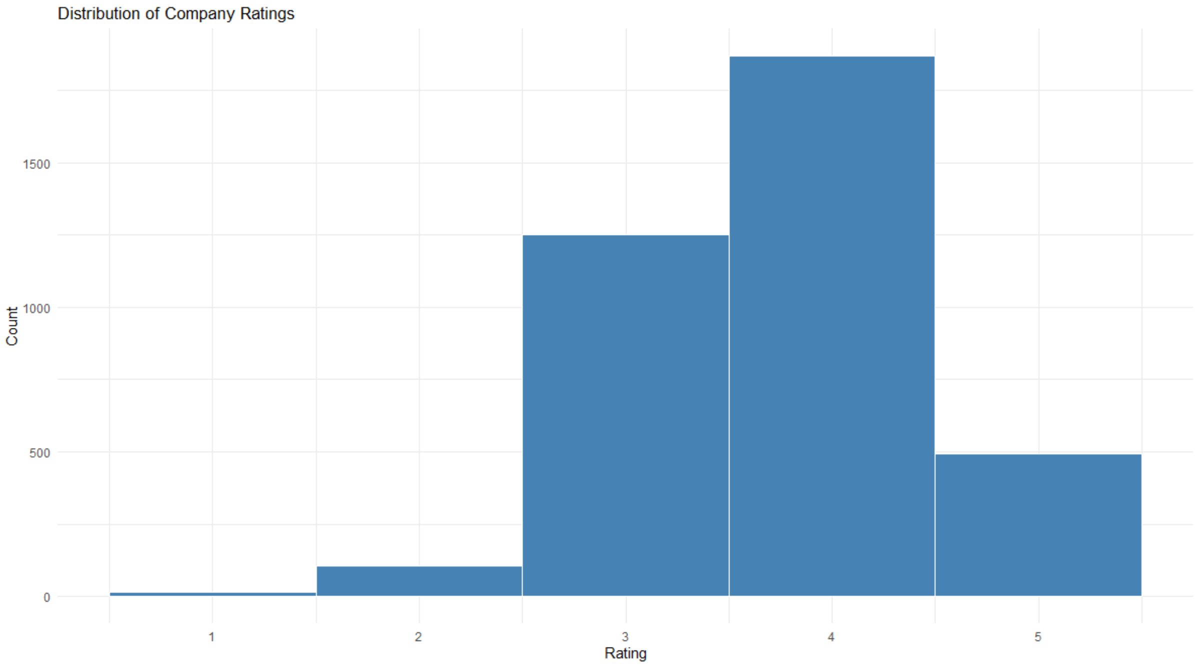

Fig. 11. Histogram of Ratings

Fig. 11 shows that Rating number 4 appears the most, followed 3, then 5. The least appeared Rating is 1.

Two-variables

Rating (numeric) vs Founded (numeric)

```
numeric_vars <- df %>%
  select(Rating, Founded) %>%
  na.omit()
cor_matrix <- cor(numeric_vars)
corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black",  addCoef.col =
"red")
```

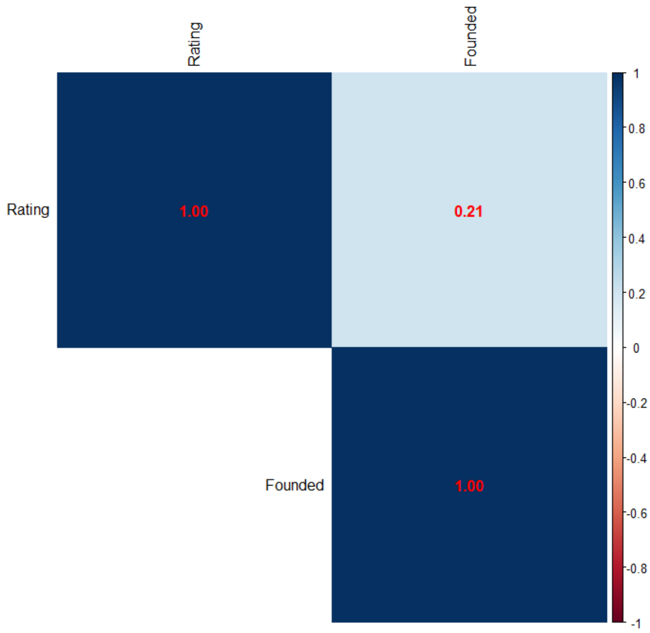

Fig. 12. Correlation between two numeric: Rating and Founded

The value of 0.21 is a weak positive linearity. This is not strong enough to suggest that there is a linear relationship, meaning that Rating increases with the year the business is founded. This is not the case.

```
# 4 Display Job Title, Rating, Location, Industry

# Display Job Title, Rating, Location, Industry

df_selected <- df %>%

  select(`Job Title`, Rating, Location, Industry)

# View the first few rows

View(df_selected)
```

	Job Title	Rating	Location	Industry
1	Business Analyst - Clinical & Logistics Platform	3.9	New York, NY	Health Care Services & Hospitals
2	Business Analyst	3.8	New York, NY	Venture Capital & Private Equity
3	Data Analyst	3.6	Florham Park, NJ	Biotech & Pharmaceuticals
4	Information Security Analyst, Incident Response	3.6	Franklin Lakes, NJ	Health Care Products Manufacturing
5	Analyst - FP&A Global Revenue	3.4	New York, NY	Internet
6	Data Analyst	3.4	Lyndhurst, NJ	Internet
7	Investment Analyst - Graduate	3.8	New York, NY	Insurance Agencies & Brokerages
8	IT Business Process Analysis	3.8	Jersey City, NJ	Biotech & Pharmaceuticals
9	Tolling Business Analyst	4.2	New York, NY	Architectural & Engineering Services
10	Business Analyst - Risk	4.4	New York, NY	Consulting
11	PGIM Investments - Research Analyst, Strategic Investment R...	3.9	Newark, NJ	Investment Banking & Asset Management
12	Business Development Analyst	4.4	New York, NY	Computer Hardware & Software
13	Senior Analyst, Business Intelligence (Looker BI)	4.1	New York, NY	Enterprise Software & Network Solutions
14	Business Systems Analyst	3.6	Brooklyn, NY	K-12 Education
15	Business Analyst	4.2	New York, NY	Consulting
16	Digital Analyst	3.1	Hackensack, NJ	Food & Beverage Manufacturing
17	Municipal Structured Products Analyst	3.7	New York, NY	Investment Banking & Asset Management
18	Senior Business Analyst	3.8	New York, NY	Health Care Services & Hospitals
19	Business Analyst	3.8	New York, NY	Consulting
20	Business Intelligence Analyst	3.4	New York, NY	Financial Transaction Processing
21	Business Analyst - CRM	3.6	New York, NY	Investment Banking & Asset Management
22	Senior FP&A Analyst	4.2	New York, NY	Lending
23	Business Analyst	3.8	New York, NY	TV Broadcast & Cable Networks

Fig. 13. Displaying the three selected fields

```
# 5a. Top 20 Industries (must be unique values, no duplicates)
top_20_industries <- df %>%
  filter(!is.na(Industry)) %>%             # Industry is not NA
  distinct() %>%                           # make it unique or distinct
  count(Industry, sort = TRUE) %>%
  slice_max(n, n = 20)                     # top 20
View(top_20_industries)
```

	Industry	n
1	IT Services	644
2	Staffing & Outsourcing	509
3	Computer Hardware & Software	243
4	Consulting	233
5	Health Care Services & Hospitals	167
6	Insurance Carriers	163
7	Investment Banking & Asset Management	157
8	Enterprise Software & Network Solutions	123
9	Banks & Credit Unions	109
10	Internet	79
11	Accounting	78
12	Advertising & Marketing	76
13	Aerospace & Defense	72
14	Biotech & Pharmaceuticals	63
15	Federal Agencies	56
16	Colleges & Universities	53
17	Lending	41
18	Energy	35

Fig. 14. Top 20 Industries (unique and no duplicates)

```
# 5b. Top 20 Sectors (must be unique, no duplicates)
top_20_sectors <- df %>%
  filter(!is.na(Sector)) %>%  # Sector not NA
  distinct() %>%
  count(Sector, sort = TRUE) %>%
  slice_max(n, n=20)
View(top_20_sectors)
```

Filter

	Sector	n
1	Information Technology	1089
2	Business Services	888
3	Finance	361
4	Insurance	178
5	Health Care	167
6	Manufacturing	131
7	Government	85
8	Accounting & Legal	82
9	Aerospace & Defense	72
10	Education	72
11	Retail	71
12	Biotech & Pharmaceuticals	63
13	Oil, Gas, Energy & Utilities	61
14	Media	38
15	Transportation & Logistics	35
16	Non-Profit	29
17	Consumer Services	25
18	Construction, Repair & Maintenance	21

Fig. 15. Top 20 Sectors

```
# 5c. Top 20 Headquarters (must be unique, no duplicates)

top_20_headquarters <- df %>%

 filter(!is.na(Headquarters)) %>%                        # Headquarters not NA

 distinct() %>%

 count(Headquarters, sort = TRUE) %>%                     # descending

 slice_max(n, n=20)                                       # top 20


View(top_20_headquarters)
```

	Headquarters	n
1	New York, NY	242
2	San Diego, CA	186
3	Woodridge, IL	178
4	Chicago, IL	127
5	Houston, TX	119
6	Jacksonville, FL	86
7	San Antonio, TX	69
8	Dallas, TX	64
9	Austin, TX	62
10	Irving, TX	61
11	Los Angeles, CA	59
12	Atlanta, GA	58
13	Tampa, FL	54
14	Philadelphia, PA	47
15	San Francisco, CA	46
16	San Jose, CA	44
17	Phoenix, AZ	41
18	Princeton, NJ	41

Fig. 16. Displaying top 20 headquarters

6a. Top 15 jobs based on Rating

The following shows the top 20 sectors

```
top_20_sectors <- df %>%
 filter(!is.na(Sector)) %>%  # Sector not NA
 distinct() %>%
 count(Sector, sort = TRUE) %>%
 slice_max(n, n=20) %>%
 View()
```

	Job_Category	avg_rating
1	Business Analyst	3.85
2	Data Analyst	3.83
3	ERP Analyst	3.81
4	Technical Analyst	3.78
5	Non-Analyst	3.74
6	Research Analyst	3.74
7	Operations Analyst	3.67
8	Other Analyst	3.65
9	Finance Analyst	3.63
10	BI Analyst	3.62
11	Marketing Analyst	3.61
12	Product Analyst	3.54

Fig. 17. Displaying top 12 Job Titles

```
# 6b
top_15_jobs_consulting <- df %>%
 filter(!is.na(`Job_Category`) & !is.na(Rating) & !is.na(Industry)) %>%
 filter(Industry == 'Consulting') %>%
 group_by(`Job_Category`, Industry) %>%
 summarise(avg_rating = round(mean(Rating, na.rm = TRUE), 1), .groups='drop') %>%
 ungroup() %>%
 arrange(desc(avg_rating)) %>%
 slice_head(n=15) %>%
 View()
```

	Job_Category	Industry	avg_rating
1	BI Analyst	Consulting	4.3
2	Technical Analyst	Consulting	4.0
3	Business Analyst	Consulting	3.9
4	Finance Analyst	Consulting	3.9
5	Non-Analyst	Consulting	3.9
6	Data Analyst	Consulting	3.8
7	Other Analyst	Consulting	3.8
8	Operations Analyst	Consulting	3.2
9	ERP Analyst	Consulting	1.0

Fig. 18. Displaying Top 12 Jobs based on Rating under consulting.

```
# 6c
bottom_15_jobs_rating <- df %>%
 filter(!is.na(`Job_Category`) & !is.na(Rating)) %>%
 group_by(`Job_Category`) %>%
 summarise(avg_rating = round(mean(Rating, na.rm = TRUE), 1)) %>%
 ungroup() %>%
 arrange(avg_rating) %>%
 slice_head(n=15) %>%
 View()
```

	Job_Category	avg_rating
1	Product Analyst	3.5
2	BI Analyst	3.6
3	Finance Analyst	3.6
4	Marketing Analyst	3.6
5	Non-Analyst	3.7
6	Operations Analyst	3.7
7	Other Analyst	3.7
8	Research Analyst	3.7
9	Business Analyst	3.8
10	Data Analyst	3.8
11	ERP Analyst	3.8
12	Technical Analyst	3.8

Fig. 19. Displaying top 12 Jobs based on Rating.

```
# 7a Top 10 Companies with rating greater than 3 and under industry â??Consultingâ??

top_consulting_companies <- df %>%

 filter(Industry == "Consulting" & !is.na(`Company Name`) & !is.na(Rating) & Rating > 3) %>%

 group_by(`Company Name`) %>%

 summarise(

  avg_rating = round(mean(Rating, na.rm = TRUE), 1),

  count = n()

 ) %>%

 ungroup()  %>%

 arrange(desc(avg_rating)) %>%

 slice_head(n = 10)
```

```
ggplot(top_consulting_companies, aes(x = fct_reorder(`Company Name`, avg_rating), y
= avg_rating)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Top 10 Consulting Companies (Rating > 3)",
    x = "Company Name",
    y = "Average Rating"
  ) +
  theme_minimal()
```

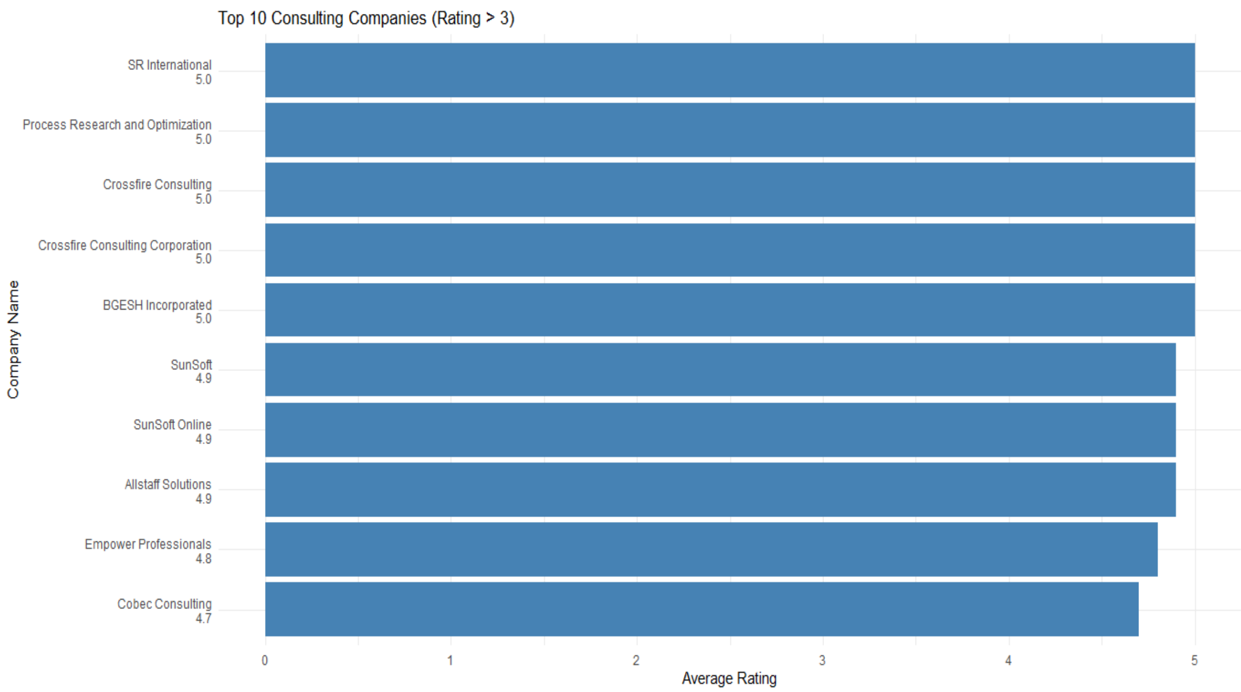

Fig. 20. Top 10 companies under consulting with rating greater than 3.

```
# 7b Top 10 Companies with rating greater than 3 and under industry Energy
top_consulting_companies_energy <- df %>%
  filter(Industry == "Energy" & !is.na(`Company Name`) & !is.na(Rating) & Rating > 3)
%>%
  group_by(`Company Name`) %>%
```

```
 summarise(
  avg_rating = round(mean(Rating, na.rm = TRUE), 1),
  count = n()
 ) %>%
 ungroup()  %>%
 arrange(desc(avg_rating)) %>%
 slice_head(n = 10)

ggplot(top_consulting_companies_energy, aes(x = fct_reorder(`Company Name`,
avg_rating), y = avg_rating)) +
 geom_col(fill = "steelblue") +
 coord_flip() +
 labs(
  title = "Top 10 Consulting Companies (Rating > 3)",
  x = "Company Name",
  y = "Average Rating"
 ) +
 theme_minimal()
```

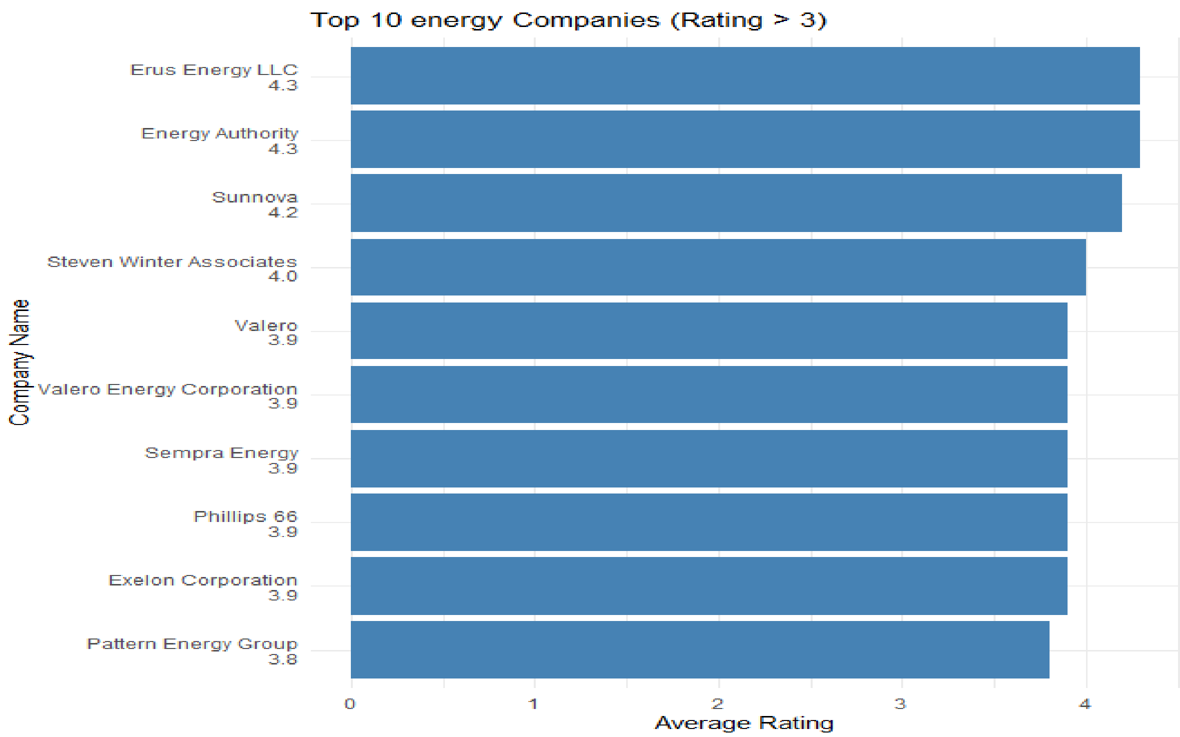

Fig. 21. Top 10 companies with rating greater than 3

```
# 7c Top 10 Companies with rating greater than 3 and under industry Accounting
top_consulting_companies_accounting <- df %>%
 filter(Industry == "Accounting" & !is.na(`Company Name`) & !is.na(Rating) & Rating > 3
 ) %>%
 group_by(`Company Name`) %>%
 summarise(
  avg_rating = round(mean(Rating, na.rm = TRUE), 1),
  count = n()
 ) %>%
 ungroup()  %>%
 arrange(desc(avg_rating)) %>%
 slice_head(n = 10)
ggplot(top_consulting_companies_accounting, aes(x = fct_reorder(`Company Name`, avg_rating), y = avg_rating)) +
 geom_col(fill = "steelblue") +
```

```
coord_flip() +
labs(
  title = "Top 10 Consulting Companies (Rating > 3)",
  x = "Company Name",
  y = "Average Rating"
) +
theme_minimal()
```

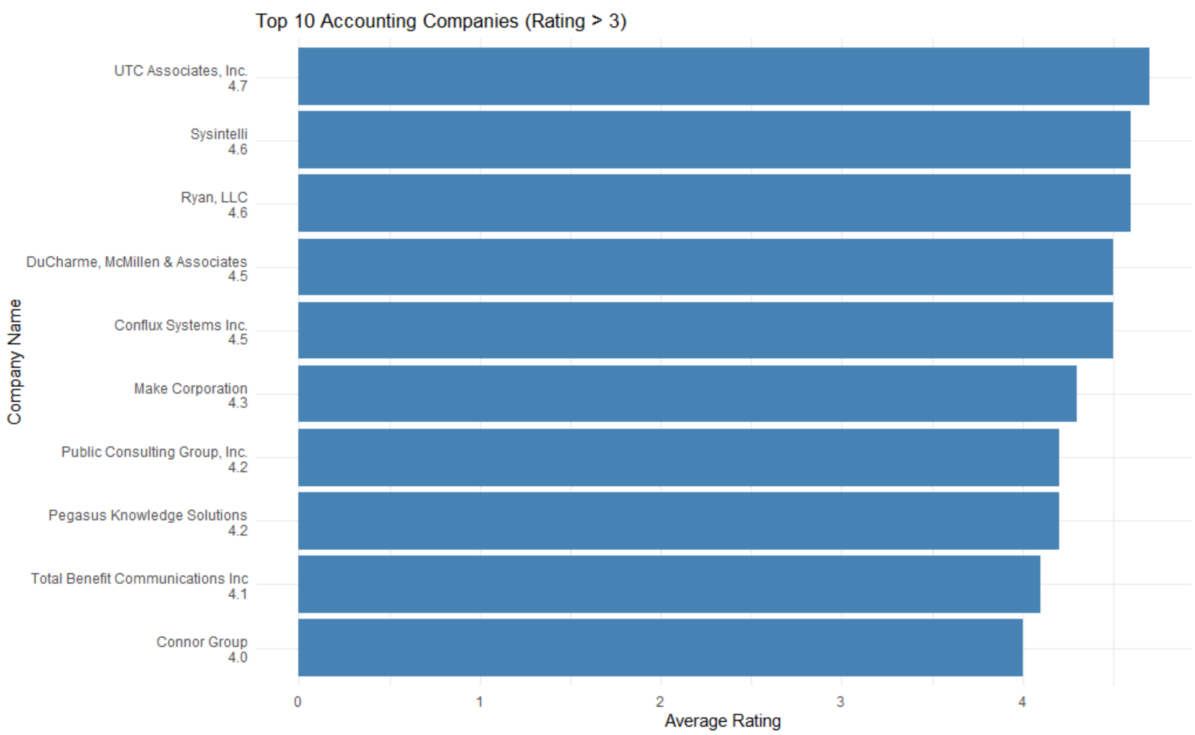

Fig. 22. Top 15 jobs with rating greater than 3 and under accounting.

Fig. 22 shows top 10 companies with a rating of greater than 3 and under accounting. The First is UTC Associates Inc., having a rating of 4.7, the last is Connor Group, with a rating 4.0.

More summaries and plots.

Top 10 Job Locations

df %>%

```
filter(!is.na(Location)) %>%
count(Location) %>%
top_n(10, n) %>%
ggplot(aes(x=fct_reorder(Location, n), y=n)) +
geom_col(fill="steelblue") + coord_flip() +
labs(
  title = "Top 10 Job Locations",
  x = "Locations",
  y = "Number of Jobs"
) +
theme_minimal()
```

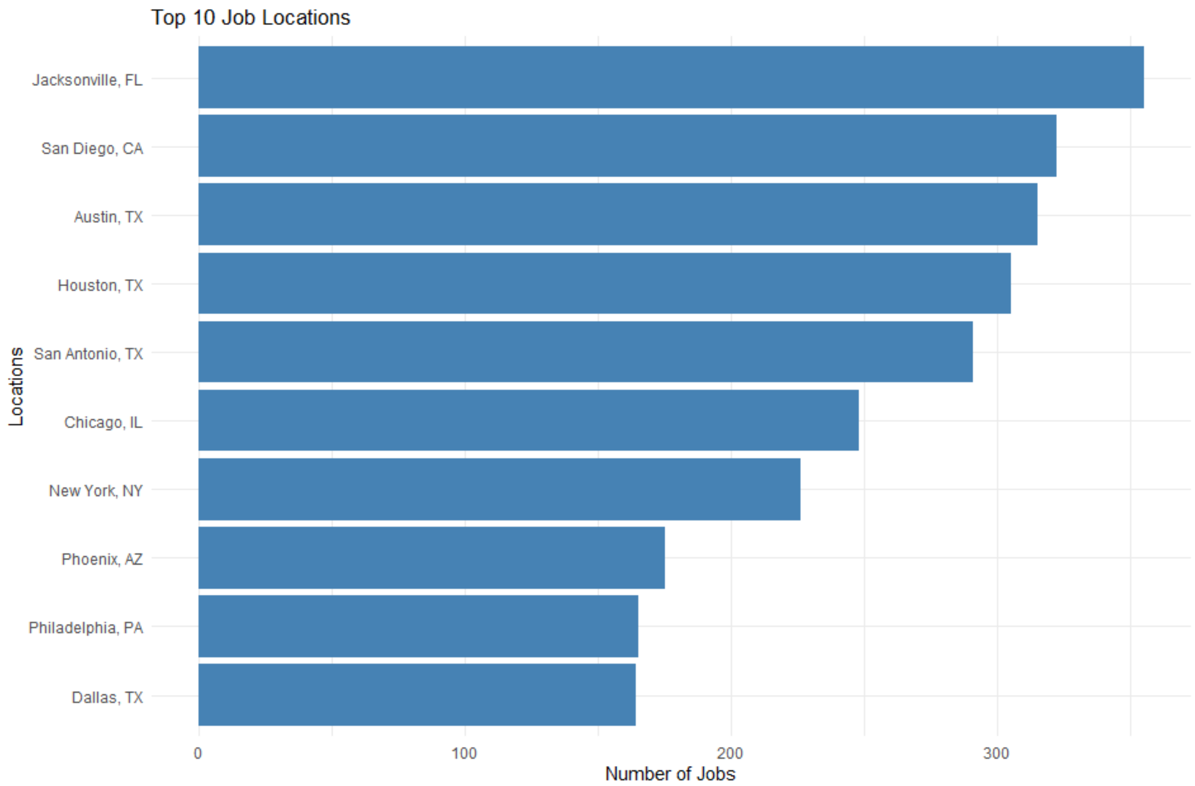

Fig. 23. Top 10 Job Locations.

The fig above (23) shows the top 10 job locations. In addition to showing the top 10 cities, I observed from the result that Texas has four of its cities in the top 10.

Top 10 Job Titles with highest average salary

```
df %>%
  mutate(avg_salary = (salary_lower + salary_upper)/2) %>%
  filter(!is.na(Job_Category) & !is.na(salary_lower) & !is.na(salary_upper) &
!is.na(avg_salary)) %>%
  group_by(Job_Category) %>%
  summarise(avg_salary = round(mean(avg_salary, na.rm = TRUE))) %>%
  ungroup() %>%
  arrange(desc(avg_salary)) %>%
  slice_head(n = 10) %>%
  View()
```

	Job_Category	avg_salary
1	Business Analyst	79128
2	Technical Analyst	79104
3	BI Analyst	78114
4	Finance Analyst	75612
5	Data Analyst	75086
6	Product Analyst	74840
7	Research Analyst	73950
8	Operations Analyst	73460
9	Other Analyst	73037
10	Non-Analyst	72323

Fig. 24. Top 10 Jobs with highest average salary

The figure above (fig. 24) highlights the top 10 jobs (Job Titles) with the highest average salary. Here we Business analyst on top, followed by Technical Analyst, then by BI Analyst, and so no. We can see the best-paying Job Titles on the average.

Top 10 industries based on Rating

```
df_clean <- df %>%
```

```
 filter(!is.na(Industry), !is.na(Rating))

# Plot boxplot
ggplot(df_clean, aes(x = fct_lump(Industry, 10), y = Rating)) +
 geom_boxplot(fill = "lightblue", outlier.color = "red") +
 coord_flip() +
 labs(title = "Rating Distribution by Industry (Top 10)",
    x = "Industry",
    y = "Rating") +
 theme_minimal()
```

Fig. 25. Rating distribution by industry

Based on the above boxplots, IT Services industry has median more than 4. 50% of the ratings falls between 5 to 3.6. Internet has median around 3.8 and 50% of the ratings is 3.7 to 4.4, which is good. Then, Consulting and Computer Hardware & Software industries have a median of 3.8 and 50% of their ratings falls between 3.5 to 4.3.

```
# Company Size Vs. Average Salary
salary_based_onSize <- df %>%
  mutate(avg_salary = (salary_lower + salary_upper)/2) %>%
  filter(!is.na(Size) & !is.na(avg_salary)) %>%
  group_by(Size) %>%
  summarise(avg_salary = round(mean(avg_salary, na.rm = TRUE), 0)) %>%
  arrange(desc(avg_salary)) %>%
  ungroup()

ggplot(salary_based_onSize, aes(x=Size, y=avg_salary, fill=Size)) +
  geom_col() +
  labs(title="Average Salary vs. Company Size", x = "Size of Company", y = "Average
Salary") +
  theme_minimal()
```

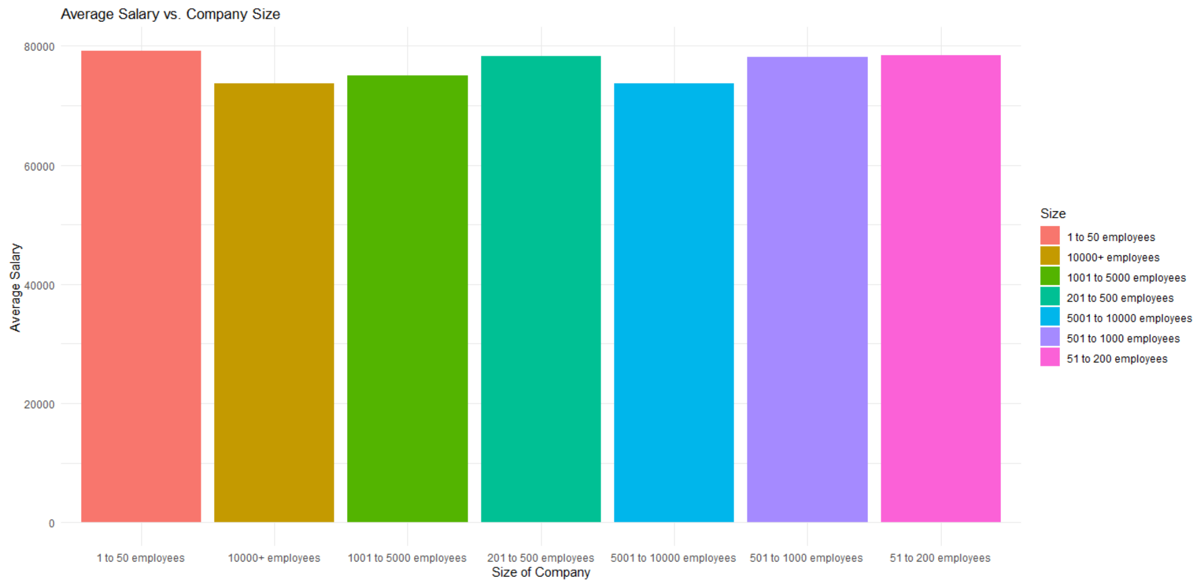

Fig. 26. Company Size Vs. Average Salary

Fig. 26 shows that company sizes: 1 to 50 employees, 51 to 200 employees, and 201 to 500 employees pay the same and highest compared to the companies with higher sizes of employees. The conclusion is that the size of the company does not affect the average salary. Larger companies do not pay better.

Top 10 Sector Using the Easy Apply

```
top_10_sectors_using_easyApply <- df %>%
  filter(!is.na(Sector) & !is.na(`Easy Apply`)) %>%
  filter(`Easy Apply` == 1) %>%
  group_by(Sector) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  slice_head(n = 10) %>%
  ungroup() %>%
  View()
```

	Sector	count
1	Business Services	51
2	Information Technology	43
3	Manufacturing	7
4	Finance	4
5	Government	4
6	Health Care	4
7	Transportation & Logistics	4
8	Oil, Gas, Energy & Utilities	3
9	Insurance	2
10	Media	2

Fig. 27. Top 10 Sectors for Easy Apply

The above figure shows top 10 sectors in the dataset that have easy apply workflow. The charts for the above summaries.

```
ggplot(top_10_sectors_using_easyApply, aes(x=reorder(Sector, count), y = count)) +
  geom_col(fill='steelblue') +
  coord_flip() +
```

```
  labs(title = "Top 10 Sectors With Easy Apply", x = "Sector", y = "Number of Easy Apply Jobs") +
  theme_minimal()
```

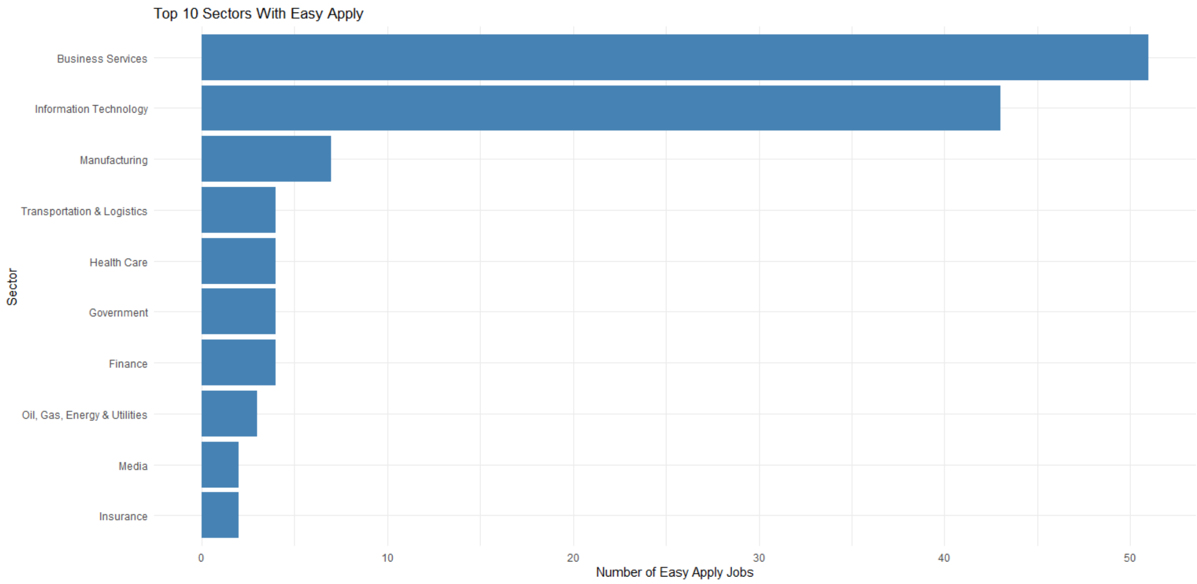

Fig. 28. The column chart of the top 10 sectors using the Easy Apply

As the summaries, the first two are Business Services and Information Technology with 51 and 43 respectively. Then, Manufacturing comes third with 7. These sectors have made their application workflow easy.

```
# Top 10 companies with highest Job Listings
top_10_companies_by_listings <- df %>%
  filter(!is.na(`Company Name`)) %>%
  group_by(`Company Name`) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  slice_head(n = 10) %>%
  ungroup() %>%
  View()
```

	Company Name	count
1	Staffigo Technical Services, LLC 5.0	178
2	Kforce 4.1	37
3	Citi 3.7	30
4	Diverse Lynx 3.9	30
5	Solekai Systems Corp 4.2	30
6	Randstad 3.6	25
7	Apex Systems 3.8	21
8	Robert Half 3.5	21
9	Lorven Technologies Inc 4.0	20
10	MUFG 3.1	20

Fig. 29. Top 10 companies with highest Job Listings.

Fig. 29 shows the top 10 companies that have the highest number of job listings. The first is Staffigo Technical Services, LLC with 178 jobs, second place is Kforce with 37, followed by Citi, Diverse Lynx, Solekai Systems Corp that have 30 job listings each.

```
# The city with the highest number of job postings
top_city_with_highest_jobs <- df %>%
 filter(!is.na(Location)) %>%
 group_by(Location) %>%
 summarise(number_of_jobs = n()) %>%
 arrange(desc(number_of_jobs)) %>%
 slice_head(n=1) %>%
 ungroup() %>%
 View()
```

	Location	number_of_jobs
1	Jacksonville, FL	355

Fig. 30. The city with the highest job listing.

The city with the highest job postings is Jacksonville, FL. It has 355 jobs posted.

```
# Proportion of jobs by type of ownership
df %>%
 filter(!is.na(`Type of ownership`)) %>%
 group_by(`Type of ownership`) %>%
 summarise(number_of_jobs = n()) %>%
 mutate(proportion = round(number_of_jobs/sum(number_of_jobs), 3)) %>%
 arrange(desc(proportion)) %>%
 ungroup() %>%
 View()
```

	Type of ownership	number_of_jobs	proportion
1	Company - Private	2167	0.585
2	Company - Public	1015	0.274
3	Subsidiary or Business Segment	184	0.050
4	Nonprofit Organization	120	0.032
5	Government	93	0.025
6	College / University	42	0.011
7	Hospital	26	0.007
8	Contract	24	0.006
9	Other Organization	13	0.004
10	Private Practice / Firm	8	0.002
11	School / School District	9	0.002
12	Self-employed	2	0.001
13	Franchise	1	0.000

Fig. 31. Proportion of jobs by type of ownership.

The above figure, fig. 31, shows the proportions of jobs for each type of ownership. Private owned companies account for 58% of the total jobs, followed by public owned companies that account for 27% of the jobs listed in the dataset. Others account for relatively insignificant amounts to the job postings.

Conclusions