ALY 6010: Probability Theory and Introductory Statistics

Milestone 1

Group 6

Poorva Joshi

V S N Sai Krishna Mohan Kocherlakota

Yash Gokhale

Rohit Lallan Gupta

Department of Project Management

College of Professional Studies

Northeastern University

Boston, MA

Prof. Yun Jiyoung

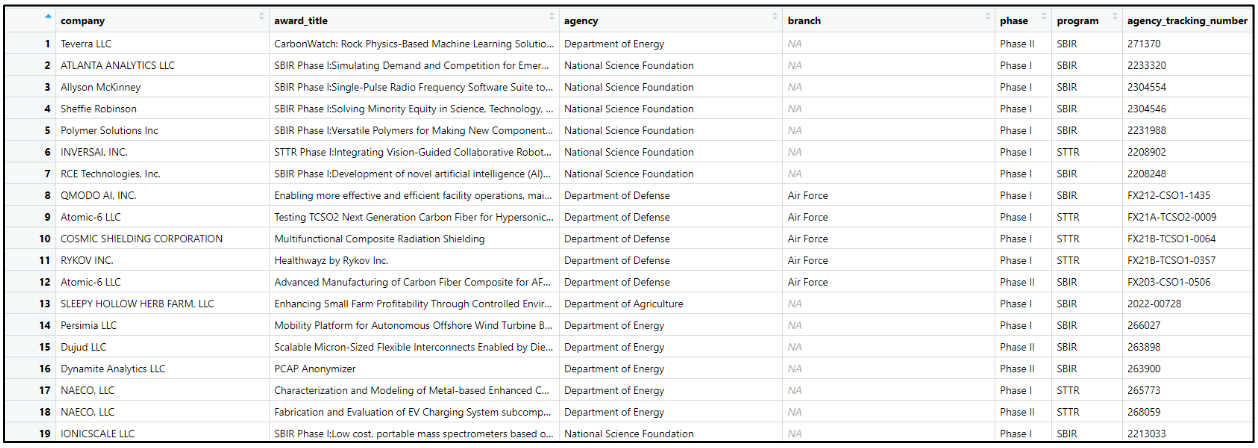
19th November 2023

Table of Contents:

INTRODUCTION

□ Summary of dataset

The dataset we used for descriptive statistics and exploratory data analysis is the US Seed Foundation dataset, which has 30 attributes that include company, agency, award title, stage, program, counties, city, zip code, contact name, contact name, contact phone number, and address1 and it shows the awards the company received in different segments covering different categories and subcategories. The collection also includes statistics on awards received by various city, state and region companies. To gain insight into this data set, we conducted an experimental data analysis on it. We also calculate descriptive statistics for the material. We developed visualizations from this data set to achieve my goal. Here we refer to a dataset from two states, Georgia and Tennessee, for this task.



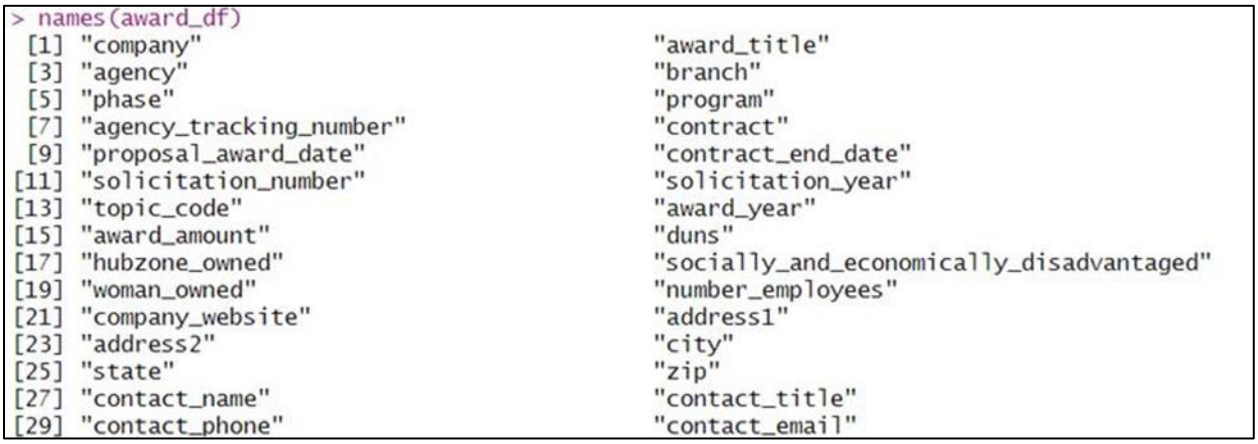
	company	award_title	agency	branch	phase	program	agency_tracking_number
1	Teverra LLC	CarbonWatch: Rock Physics-Based Machine Learning Solutio...	Department of Energy	NA	Phase II	SBIR	271370
2	ATLANTA ANALYTICS LLC	SBIR Phase I:Simulating Demand and Competition for Emer...	National Science Foundation	NA	Phase I	SBIR	2233320
3	Allyson McKinney	SBIR Phase I:Single-Pulse Radio Frequency Software Suite to...	National Science Foundation	NA	Phase I	SBIR	2304554
4	Sheffie Robinson	SBIR Phase I:Solving Minority Equity in Science, Technology, ...	National Science Foundation	NA	Phase I	SBIR	2304546
5	Polymer Solutions Inc	SBIR Phase I:Versatile Polymers for Making New Component...	National Science Foundation	NA	Phase I	SBIR	2231988
6	INVERSAI, INC.	STTR Phase I:Integrating Vision-Guided Collaborative Robot...	National Science Foundation	NA	Phase I	STTR	2208902
7	RCE Technologies, Inc.	SBIR Phase I:Development of novel artificial intelligence (AI)...	National Science Foundation	NA	Phase I	SBIR	2208248
8	QMODO AI, INC.	Enabling more effective and efficient facility operations, mai...	Department of Defense	Air Force	Phase I	SBIR	FX212-CSO1-1435
9	Atomic-6 LLC	Testing TCSO2 Next Generation Carbon Fiber for Hypersonic...	Department of Defense	Air Force	Phase I	STTR	FX21A-TCSO2-0009
10	COSMIC SHIELDING CORPORATION	Multifunctional Composite Radiation Shielding	Department of Defense	Air Force	Phase I	STTR	FX21B-TCSO1-0064
11	RYKOV INC.	Healthwayz by Rykov Inc.	Department of Defense	Air Force	Phase I	STTR	FX21B-TCSO1-0357
12	Atomic-6 LLC	Advanced Manufacturing of Carbon Fiber Composite for AF...	Department of Defense	Air Force	Phase II	SBIR	FX203-CSO1-0506
13	SLEEPY HOLLOW HERB FARM, LLC	Enhancing Small Farm Profitability Through Controlled Envir...	Department of Agriculture	NA	Phase I	SBIR	2022-00728
14	Persimia LLC	Mobility Platform for Autonomous Offshore Wind Turbine B...	Department of Energy	NA	Phase I	SBIR	266027
15	Dujud LLC	Scalable Micron-Sized Flexible Interconnects Enabled by Die...	Department of Energy	NA	Phase II	SBIR	263898
16	Dynamite Analytics LLC	PCAP Anonymizer	Department of Energy	NA	Phase II	SBIR	263900
17	NAECO, LLC	Characterization and Modeling of Metal-based Enhanced C...	Department of Energy	NA	Phase I	STTR	265773
18	NAECO, LLC	Fabrication and Evaluation of EV Charging System subcomp...	Department of Energy	NA	Phase II	STTR	268059
19	IONICSCALE LLC	SBIR Phase I:Low cost, portable mass spectrometers based o...	National Science Foundation	NA	Phase I	SBIR	2213033

In the above screenshot, it is clear that this dataset consists of 30 attributes. Here in order to know the total number of fields and records, dim() function is used which displays the total number of rows and columns.

Columns- ‘company’, ‘award title’, ‘agency’, ’branch’, ‘program’, ‘agency tracking number’, ‘contract’, ‘proposal award date’, ‘contract end date’, etc.

DATA ANALYSIS

The initial task was to bring the dataset into R Studio and do preliminary research to understand its composition and quality. This step is crucial because it sets the stage for all subsequent analyses. Post-import diagnostics indicated that the data needed to be cleaned to correct missing values in several columns.



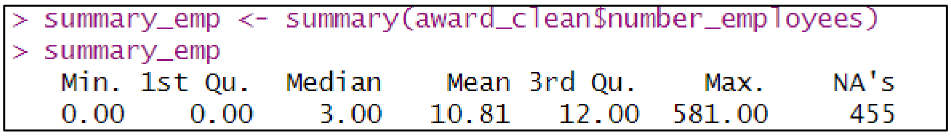
```
> names(award_df)
 [1] "company"                                  "award_title"
 [3] "agency"                                   "branch"
 [5] "phase"                                    "program"
 [7] "agency_tracking_number"                   "contract"
 [9] "proposal_award_date"                      "contract_end_date"
[11] "solicitation_number"                      "solicitation_year"
[13] "topic_code"                               "award_year"
[15] "award_amount"                             "duns"
[17] "hubzone_owned"                            "socially_and_economically_disadvantaged"
[19] "woman_owned"                              "number_employees"
[21] "company_website"                          "address1"
[23] "address2"                                 "city"
[25] "state"                                    "zip"
[27] "contact_name"                             "contact_title"
[29] "contact_phone"                            "contact_email"
```

The above image shows that this data includes 30 columns and over 800 data sets, which can be utilized for additional data analysis and descriptive analysis using clean data.

□ Descriptive statistics

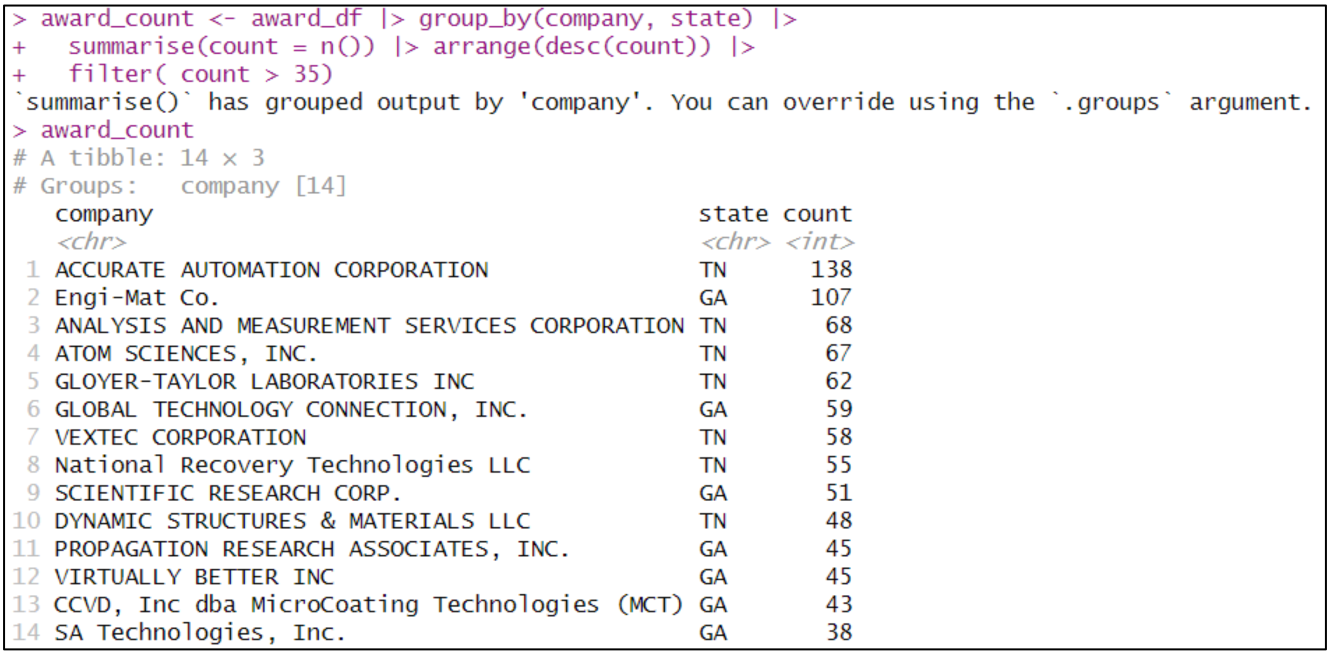
- Getting the structure of the dataset, where it is clear that the columns sales, discount and profit are of num data type, which is converted to int data type while cleaning the data.

- Descriptive summary of the dataset which displays the Min, mean, median, max.



```
> summary_emp <- summary(award_clean$number_employees)
> summary_emp
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00    0.00    3.00   10.81   12.00  581.00     455 
```

□ Analysing the Dataset



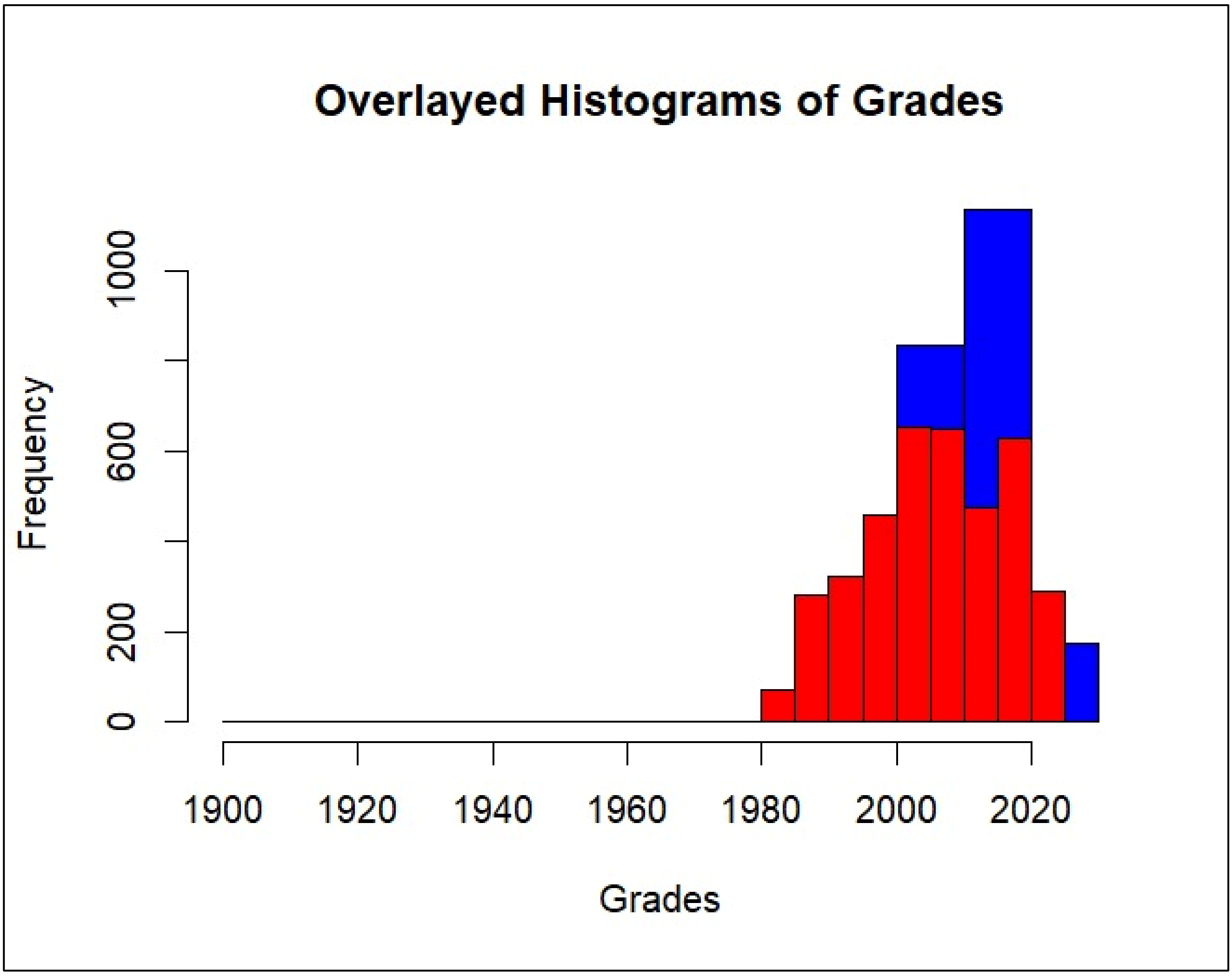
```
> award_count <- award_df |> group_by(company, state) |>
+   summarise(count = n()) |> arrange(desc(count)) |>
+   filter( count > 35)
`summarise()` has grouped output by 'company'. You can override using the `.groups` argument.
> award_count
# A tibble: 14 x 3
# Groups:   company [14]
   company                                     state count
   <chr>                                       <chr> <int>
 1 ACCURATE AUTOMATION CORPORATION             TN      138
 2 Engi-Mat Co.                                GA      107
 3 ANALYSIS AND MEASUREMENT SERVICES CORPORATION TN     68
 4 ATOM SCIENCES, INC.                         TN       67
 5 GLOYER-TAYLOR LABORATORIES INC              TN       62
 6 GLOBAL TECHNOLOGY CONNECTION, INC.          GA       59
 7 VEXTEC CORPORATION                          TN       58
 8 National Recovery Technologies LLC          TN       55
 9 SCIENTIFIC RESEARCH CORP.                   GA       51
10 DYNAMIC STRUCTURES & MATERIALS LLC          TN       48
11 PROPAGATION RESEARCH ASSOCIATES, INC.       GA       45
12 VIRTUALLY BETTER INC                        GA       45
13 CCVD, Inc dba MicroCoating Technologies (MCT) GA     43
14 SA Technologies, Inc.                       GA       38
```

The above image shows the names of the companies obtained by running a specific R script, that also displays the US state in which the company is located and the number of awards won by that respective company.

□ Data visualization

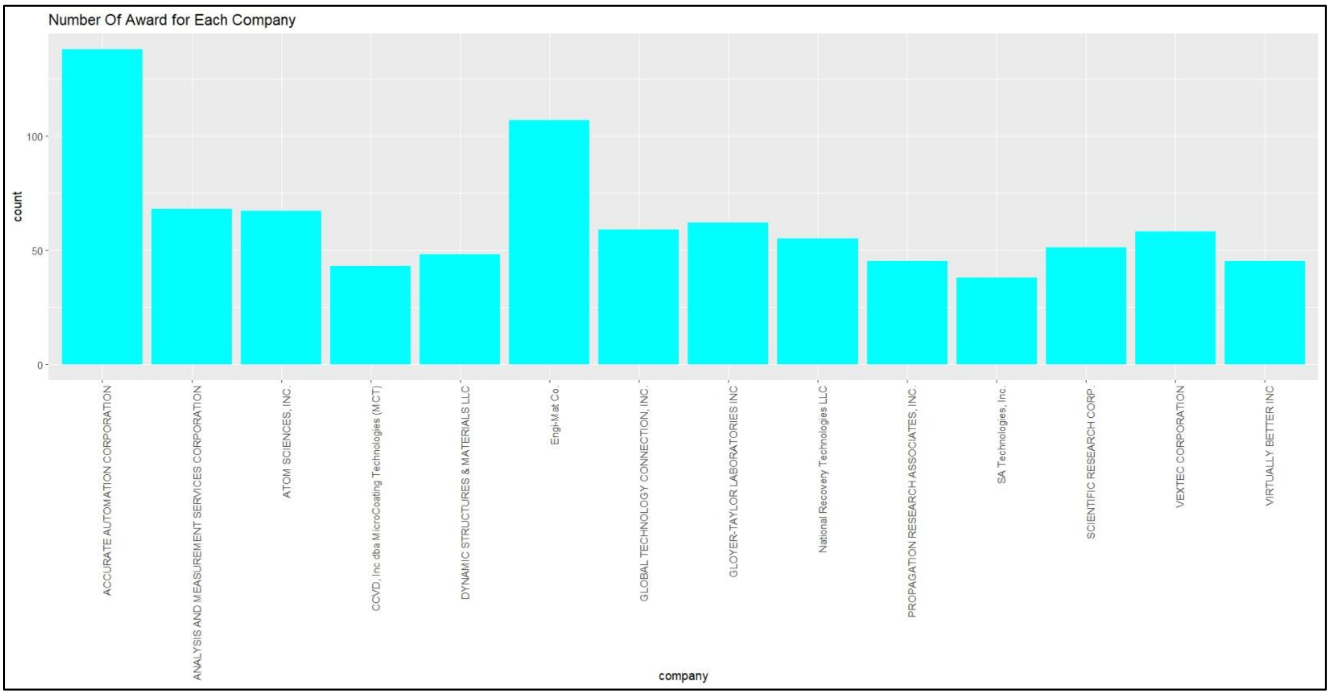
- Histograms

Overlapping histograms visually compare multiple datasets by displaying their distributions simultaneously on a single plot. They allow easy identification of similarities, differences, and potential patterns between datasets by showing overlap and distinct areas. Using different colors or transparency helps distinguish datasets, aiding in quick visual analysis of distribution characteristics.



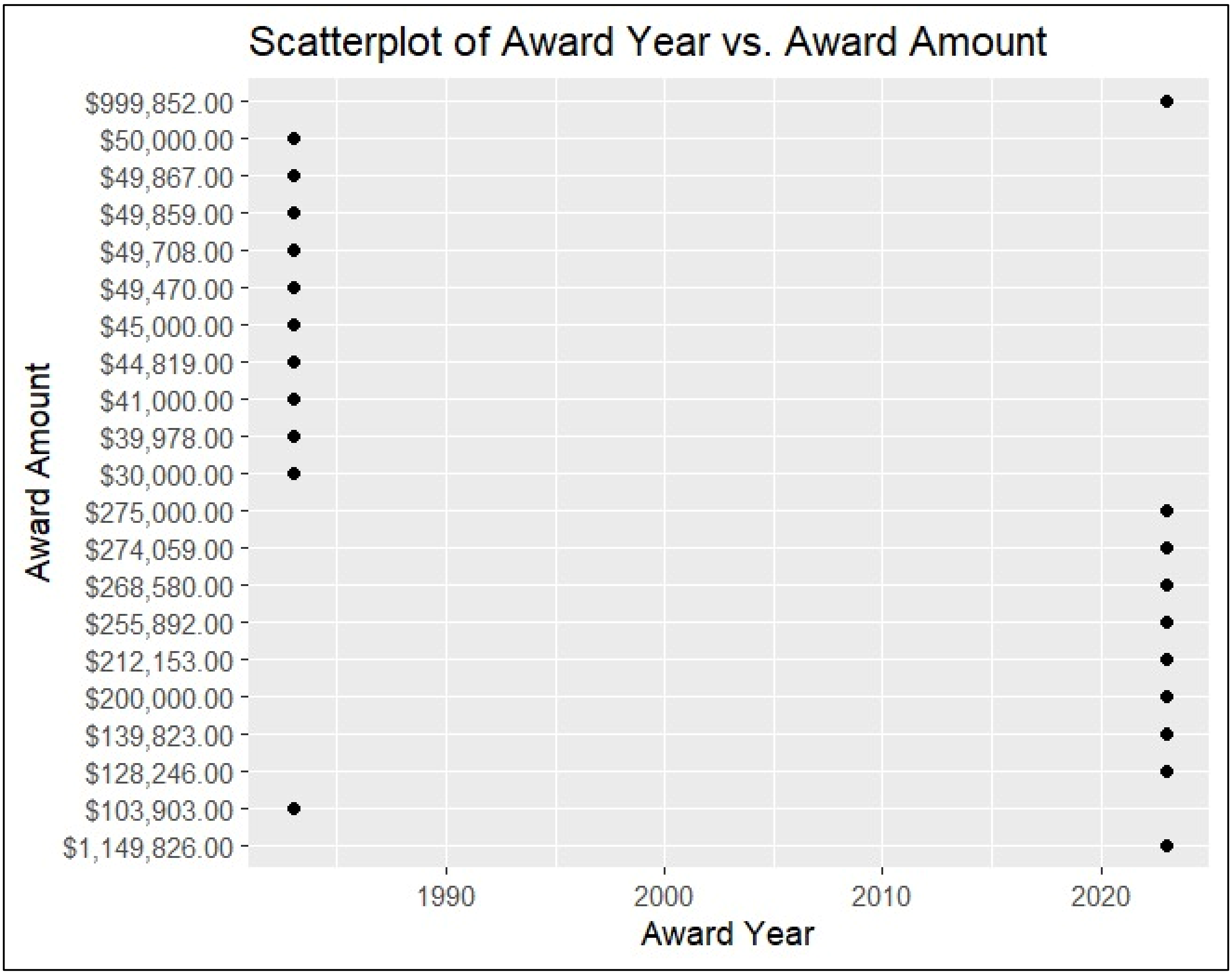
- Bar plot:

This bar chart shows the number of awards received by company and company over the years and it can be observed that the blue bar represents companies like SA Technologies Inc. The awards won by Accurate Automation Corporation are the highest.



- Scatter Plot:

Below is the scatter plot of the years 1983 and 2023. As the image shows, there is a huge difference in the award amounts in the span of 40 years.



CITATIONS

1. Zach. (2022, April 13). How to perform exploratory data analysis in R (with example). Statology. Retrieved November 11, 2023, from https://www.statology.org/exploratorydata-analysis-in-r/
2. Creating. Creating and updating figures in R. (n.d.). Retrieved November 11, 2022, from https://plotly.com/r/creating-and-updating-figures/
3. Holtz, Y. (n.d.). Density chart: The R Graph Gallery. Density Chart | the R Graph Gallery. Retrieved November 11, 2023, from https://r-graph-gallery.com/densityplot.html