ALY 6010: Probability Theory and Introductory Statistics

Final Project: Milestone 2

Group 6

Poorva Joshi

V S N Sai Krishna Mohan Kocherlakota

Yash Gokhale

Rohit Lallan Gupta

College of Professional Studies

Northeastern University

Boston, MA

Prof. Yun Jiyoung

5th December 2023

Introduction

In this project, our team aimed to conduct an exploratory data analysis and hypothesis testing on a government awards dataset. The collection includes data on a variety of awards, such as program kinds, corporate information, and award amounts. Finding relevant insights by examining possible patterns and links in the data was the main goal.

Key Findings

Business Questions: -

1. Which Companies Have Received the Most Awards?

Code:

```
> award_count <- award_df |> group_by(company, state) |>
+   summarise(count = n()) |> arrange(desc(count)) |>
+   filter( count > 35)
```

Output:

```
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
```

Hypothesis: There is a significant difference in the number of awards received by companies, and some companies are consistently awarded more frequently than others.

2. Is There a Significant Difference in Award Amounts for Socially and Economically Disadvantaged Companies?

Code:

```
> t_test_result1 <- t.test(disadvantaged, non_disadvantaged)
```

Output:

```
> t_test_result1

        Welch Two Sample t-test

data:  disadvantaged and non_disadvantaged
t = 0.55906, df = 284.33, p-value = 0.5766
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -41765.62  74902.32
sample estimates:
mean of x mean of y
 360331.1  343762.8
```

Hypothesis: There is a significant difference in the mean award amounts between socially and economically disadvantaged companies (Y) and non-disadvantaged companies (N).

3. How Does the Number of Employees Vary Across Different States?

Code:

```
> t_test_result3 <- t.test(number_employees ~ state, data = award_clean)
```

Output:

```
> summary(t_test_result3)
            Length Class  Mode
statistic   1      -none- numeric
parameter   1      -none- numeric
p.value     1      -none- numeric
conf.int    2      -none- numeric
estimate    2      -none- numeric
null.value  1      -none- numeric
stderr      1      -none- numeric
alternative 1      -none- character
method      1      -none- character
data.name   1      -none- character
```

Hypothesis: The mean number of employees varies significantly among different states, suggesting that the state of operation has an impact on the size of companies receiving awards.

Visualizations: -

1. **BarPlot-**
   The graphic shows how the average number of employees in each state compares to one another. It assists in determining which states have higher or lower average employment rates among the award-winning enterprises.

Average Number of Employees by State
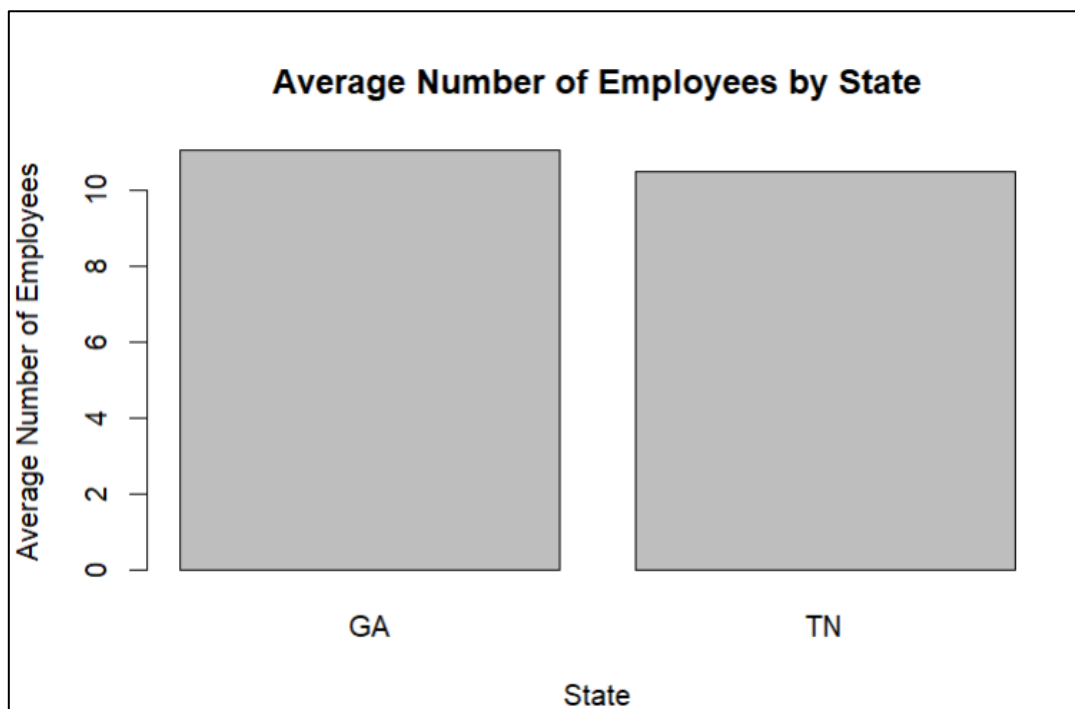
2. BoxPlot-

The box plot provides information about the distribution of the number of employees per state, including the central tendency, spread, and any outliers. It makes it possible to compare the features of workforce size visually between states.

Number of Employees by State
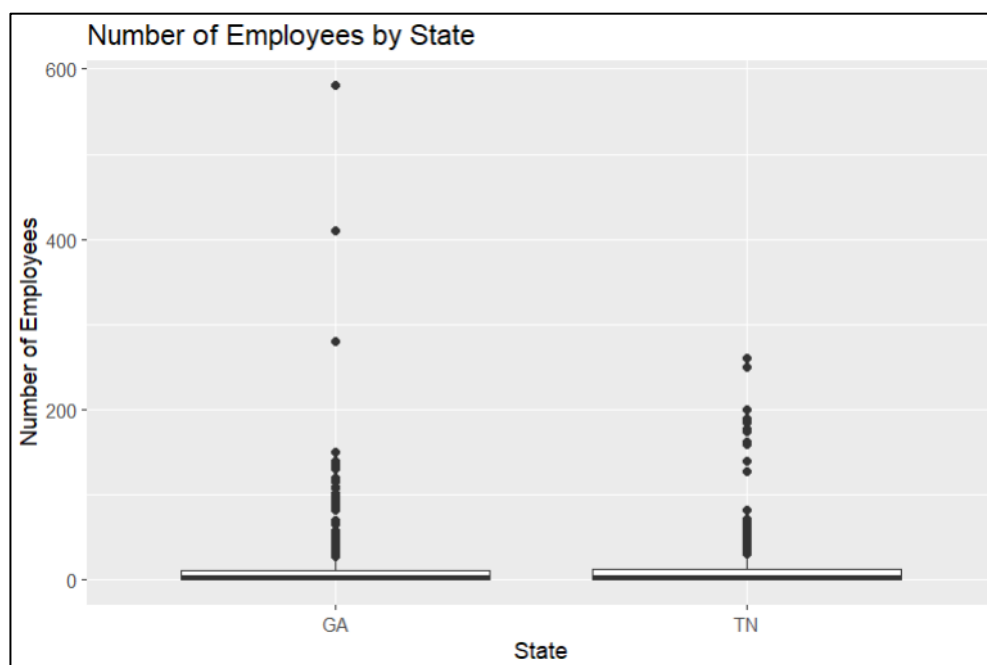
Conclusion

Our investigation offers insightful information about the variables affecting the size of government awards. The findings demonstrate how program kinds, regional location, and social and economic deprivation affect parameters connected to awards. Policymakers and other interested parties can use this information to help them make well-informed decisions about government grants.

Citations

1. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test
2. https://www.datacamp.com/tutorial/t-tests-r-tutorial