US Bachelor's Graduates: A Comprehensive R analysis on Employment Trends Upon Graduation

Belen Cerrutti

2024-03-25

Introduction

The purpose of this analysis is to understand what factors influence recent graduates' obtainment of a job and their starting salary. By analyzing patterns and correlations within the data, I aim to determine the impact of factors like GPA, gender, major and years of experience when it comes to landing a job. The goal is to help prospective students make informed decisions about their education and career paths. This study utilizes a dataset from Kaggle, specifically the "Job Placement Dataset" which can be found at this link. The dataset includes variables such as degree major, work experience, gender, and placement status, among others, offering a comprehensive overview of the job placement scenario for graduates.

Data dictionary

- *Placement status* - Being hired or not (placed = getting hired, not placed = not getting hired)
- -*Stream* - Major in College
- -*Degree* - All the students in this dataset received a Bachelor's degree

Data Loading and Preparation

Now we'll load the data and prepare it for analysis.

```
options(repos = c(CRAN = "https://cran.rstudio.com/"))
library("readxl")

df <-
read_excel("C:\\Users\\belen\\OneDrive\\PROJECTS\\R\\job_placement.xlsx")
head(df)

## # A tibble: 6 × 11
##      id name     gender   age degree stream college_name placement_status
salary
##   <dbl> <chr>    <chr>  <dbl> <chr>  <chr>  <chr>        <chr>
```

```
<dbl>
## 1     1 John Doe Male      25 Bache… Compu… Harvard Uni… Placed
60000
## 2     2 Jane Sm… Female    24 Bache… Elect… Massachuset… Placed
65000
## 3     3 Michael… Male      26 Bache… Mecha… Stanford Un… Placed
58000
## 4     4 Emily D… Female    23 Bache… Infor… Yale Univer… Not Placed
0
## 5     5 David B… Male      24 Bache… Compu… Princeton U… Placed
62000
## 6     6 Sarah W… Female    25 Bache… Elect… Columbia Un… Placed
63000
## # ℹ 2 more variables: gpa <dbl>, years_of_experience <dbl>
```

Data Exploration

Q1: Is there a significant difference in GPA between those who were hired and those who were not?

Note: In this dataset hired is referred to as 'placed' and nor hired is referred to as 'not placed'.

First, we extract GPAs for placed and not placed individuals.

```
placed_gpa <- df[df$placement_status == 'Placed', 'gpa']
not_placed_gpa <- df[df$placement_status == 'Not Placed', 'gpa']
```

We will perform a T-test to assess whether there is a significant difference in the mean GPAs of placed and not placed students. Our null hypothesis (H0) is that the mean GPA of those placed is less than or equal to those not placed. The alternative hypothesis (Ha) is that the mean GPA of those placed is greater than those not placed.

```
t_test <- t.test(placed_gpa, not_placed_gpa, paired=FALSE,
alternative="greater", conf.level = 0.95)
t_test

##
##  Welch Two Sample t-test
##
## data:  placed_gpa and not_placed_gpa
## t = 4.4426, df = 168.6, p-value = 8.019e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.03709512        Inf
## sample estimates:
## mean of x mean of y
##  3.761404  3.702308
```

Interpretation: Because the p-value is smaller than alpha (0.05), we can conclude that there is enough statistical evidence to suggest that on average, students who got a job after graduation had a higher GPA score than those students who did not land a job.

Q2: What majors result in the highest average salaries?

To answer this question, we'll look at the unique streams (majors) within the dataset and then visualize the distribution of salaries for each stream with a box plot.

```
library(stringr)
unique(df$stream)

## [1] "Computer Science"                "Electrical Engineering"
## [3] "Mechanical Engineering"          "Information Technology"
## [5] "Electronics and Communication"

df$stream <- as.factor(df$stream) #converts stream into a factor
```

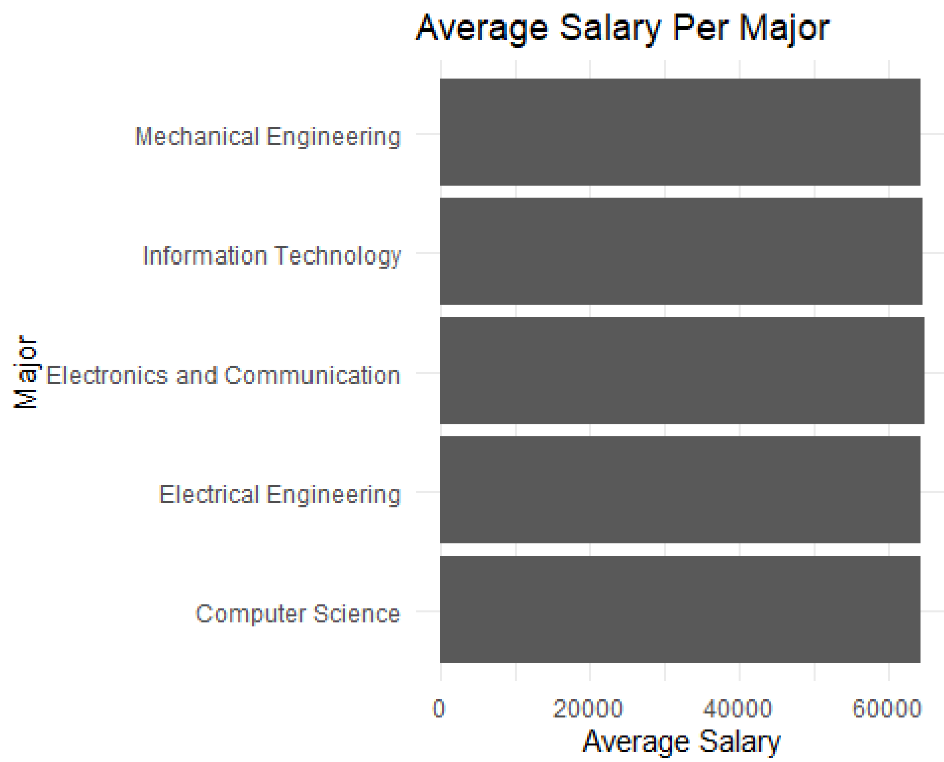
First we calculate the average salary for each major and then we visualize the data using a column chart.

```
library(dplyr)
library(ggplot2)

result <- df %>%
  filter(salary != 0) %>%
  group_by(stream) %>%
  summarise(average_salary = mean(salary, na.rm = TRUE))
result

## # A tibble: 5 × 2
##   stream                        average_salary
##   <fct>                                  <dbl>
## 1 Computer Science                      64280.
## 2 Electrical Engineering                64144.
## 3 Electronics and Communication         64832.
## 4 Information Technology                64609.
## 5 Mechanical Engineering                64356.

ggplot(data = result, aes(x = stream, y = average_salary)) +
  geom_col() +
  coord_flip() +  # This flips the x and y axes
  labs(title = "Average Salary Per Major", x = "Major", y = "Average Salary")
+
  theme_minimal()
```



Interpretation: Judging by the chart, we can see that there is practically no difference in the average salary among different majors.

Q3:Is being hired dependent on gender?

For this question, our null and alternate hypothesis are as follows:

-H0: getting hired is independent of gender

-Ha: getting hired is not independent of gender

To test our hypothesis, we will conduct a Chi-Square Test of Independence with a 5% significance level. First we compute a cross-tabulation of data

```
crosstab<-table(df$gender,df$placement_status)
crosstab

##
##          Not Placed Placed
##   Female         63    303
##   Male           67    267
```

Now we calculate probabilities.

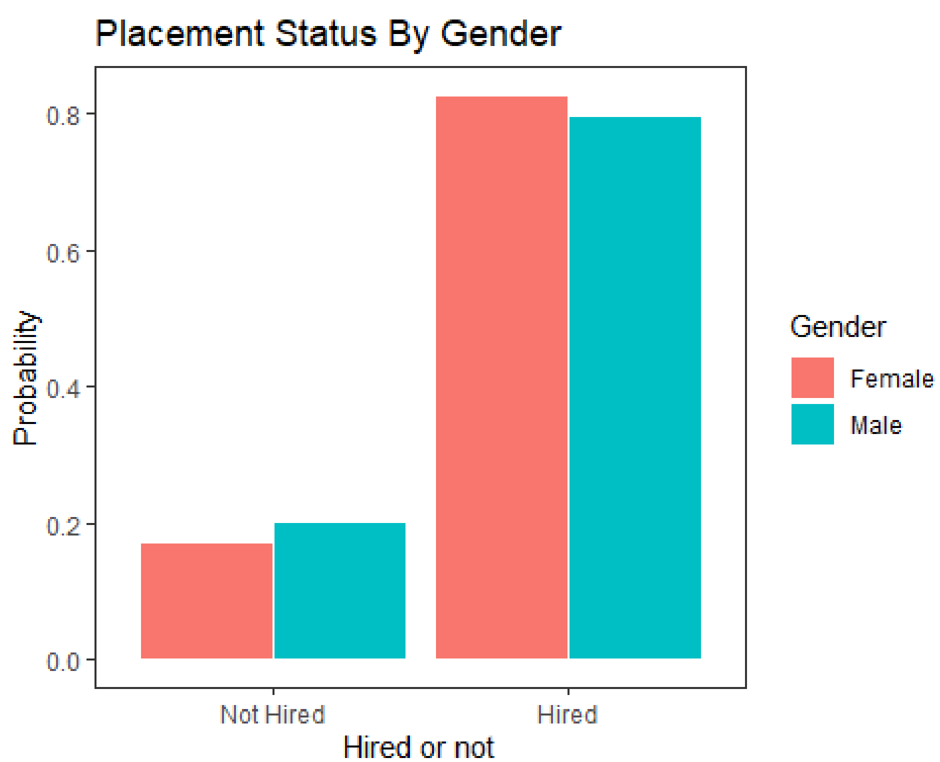
```
prob_data <- prop.table(crosstab, margin =1 ) #calculates probability of each
response
prob_data<-data.frame(prob_data) #puts probabilities in a table format
```

```r
colnames(prob_data)<- c("Gender", "Placement.status", "Probability")
prob_data #outputs probabilities
```

```
##   Gender Placement.status Probability
## 1 Female       Not Placed   0.1721311
## 2   Male       Not Placed   0.2005988
## 3 Female           Placed   0.8278689
## 4   Male           Placed   0.7994012
```

Now we visualize the data using a column chart.

```r
ggplot(prob_data, aes(x= Placement.status, y=Probability, fill=Gender))+
geom_bar(stat="identity", position="dodge", color="white")+
  labs(title ="Placement Status By Gender", y="Probability", x="Hired or
not", fill="Gender")+
  scale_x_discrete(labels = c("Not Hired", "Hired")) +
  theme_bw()+
  theme(panel.grid=element_blank())
```



Interpretation: It appears that upon graduation, women get hired at a slightly higher rate than men. While male undergraduates find a job 80% of the time, female undergraduates get hired 83% of the time, suggesting a 3% difference.

Now we will compute a Chi-square Test to determine independence

```
chi_test<-chisq.test(crosstab)
chi_test

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  crosstab
## X-squared = 0.75708, df = 1, p-value = 0.3842
```

Interpretation: Given that p-value(0.38) is bigger than alpha(0.05) we cannot reject H0. This means that there is not enough statistical evidence to conclude that getting hired is dependent on gender.

Q4: Is there any significant difference in the average salary based on the years of experience (YOE)?

To find out, we will conduct an ANOVA test with a 0.05 level of significance. The population means are as follows:

-u1=mean salary for graduates with 1 year of experience

-u2=mean salary for graduates with 2 year of experience

-u3=mean salary for graduates with 3 year of experience

Our null hypothesis (H0) is that the mean salary will be the same for 1,2 and 3 years of experience, while the alternative hypothesis (Ha) is that not all mean salaries are equal.

-H0:u1= u2 = u3

-Ha:u1≠ u2 ≠ u3

First, we will prepare the data by removing remove rows with salary of "$0".

```
df$salary <- as.numeric(df$ salary) #setting salary as numeric
ANOVA_data <- df[df$salary > 0, ] #creates a new table without salaries=0
```

Let's see the levels in years of experience.

```
ANOVA_data$years_of_experience <- factor(ANOVA_data$years_of_experience)
levels(ANOVA_data$years_of_experience)

## [1] "1" "2" "3"
```

To visualize the data we will create a box plot showing salary based on years of experience.

```
ANOVA_data %>%
  drop_na(years_of_experience)%>%
  ggplot(aes(years_of_experience,salary))+
  geom_boxplot()+
  labs(title ="Salary By Year of Experience", y="Salary", x="Years of
Experience")+
```

```r
  theme_bw()+
  theme(panel.grid= element_blank())+
  stat_summary(fun = median, geom = "text", aes(label = ..y..), vjust = -0.5,
size = 3) +
  theme(axis.text.x = element_text(size = 8)) +
  scale_x_discrete(labels = function(x) str_wrap(x, width = 10))
```

Observations:

-<u>Outliers</u>: 3 years of experience has 2 outliers.

-<u>Between group variation</u>: Low between group variation.

-<u>Within group variation:</u> Low within group variation. Year 2 is the only one with relatively high variation.

Now we will check the ANOVA assumptions for test validity.

1)Check the homogeneity of variance assumption (used Levene Test due to presence of outliers).

```r
install.packages("car")
library(car)

leveneTest(salary ~ years_of_experience, data = ANOVA_data)

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
```

```
## group   2   10.72 2.693e-05 ***
##       566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the data does not meet one of the necessary conditions for performing a traditional ANOVA, we will perform a Kruskal-Wallis rank sum test (Non-parametric alternative to one-way ANOVA test).

```
kruskal.test(salary ~ years_of_experience, data = ANOVA_data)

##
##  Kruskal-Wallis rank sum test
##
## data:  salary by years_of_experience
## Kruskal-Wallis chi-squared = 192.15, df = 2, p-value < 2.2e-16
```

Interpretation: Given the significance of the p-value (0.00) compared to a 0.05 significance level, we reject the null hypothesis and conclude that there is a statistically significant difference in the median salary across different years of experience groups.

- Pairwise Comparison:

We will perform a pairwise t-test using the Bonferroni adjustment for multiple comparisons on the `salary` variable across different `years_of_experience` levels.

```
pairwise.t.test(ANOVA_data$salary, ANOVA_data$years_of_experience,
p.adj="bonferroni")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ANOVA_data$salary and ANOVA_data$years_of_experience
##
##   1      2
## 2 0.34   -
## 3 <2e-16 <2e-16
##
## P value adjustment method: bonferroni
```

1) **1YOE VS 2 YOE**

H0: u1 = u2

Ha: u1 ≠ u2

Interpretation: given p-value(0.00)> alpha(0.05) we cannot reject H0 meaning that there is not a significant difference between mean salary with 1 and 2 YOE.

2) **1YOE VS 3 YOE**

H0: u1 = u3

Ha: u1 ≠ u3

Interpretation: given p-value(0.00)> alpha(0.05) we reject H0 meaning that there is a significant difference between mean salary with 1 and 3 YOE.

3) **2YOE VS 3 YOE**

H0: u2 = u3

Ha: u2 ≠ u3

Interpretation: given pvalue(0.00)> alpha(0.05) we reject H0 meaning that there is a significant difference between mean salary with 2 and 3 YOE.

Findings

Key insights from this analysis indicate that employed graduates generally had higher GPAs compared to their unemployed peers. Employment appears gender-neutral, signaling an equitable job market.Furthermore, salary assessments across various majors reveal a consistent average of approximately $64,000 in the Engineering and Technology sectors, underscoring the field's uniform financial prospects post-graduation. Notably, while salary differences are minimal between one to two years of experience, surpassing three years marks a notable increase in earnings.

Given these findings, students in Engineering and Technology should pursue their chosen major with enthusiasm, maintaining a strong academic record to bolster their employment prospects. While GPA is not a guaranteed predictor of job placement, it is observed that hired individuals often showcase superior academic achievements. Gaining practical experience through internships or related extracurricular activities is also recommended to enhance employability and expedite the hiring process. Finally, women contemplating a STEM career should remain confident, as the data reflects no gender-based hiring bias.