Pricilla Nakyazze Project Proposal

2025-05-04

Abstract

What is the relationship between academic performance and career success?

This study investigates the multifaceted relationship between academic performance and career success, aiming to predict job outcomes based on educational and personal attributes. With the rise of data-driven decision-making, understanding which factors most strongly correlate with professional advancement has become essential. This project focuses on evaluating key indicators such as GPA, SAT scores, networking, gender, age, and education level to assess their impact on career Job offers.

Using statistical analysis, we analyze how these variables interact to influence job placement and earnings. A particular emphasis is placed on GPA and networking_Score as potential accelerators of job placement. We aim to determine whether high academic scores alone are sufficient predictors of success, or if Age and Gender play a more significant role.

Preliminary findings suggest that while GPA and SAT scores positively correlate with initial job placement, factors like networking have a stronger influence on career development and salary trajectory. Additionally, demographic factors such as age and gender may introduce variability in outcomes, necessitating a nuanced approach to equitable career guidance.

This Project contributes to exploring and identifying the most impactful factors in bridging the gap between academic achievement and a career. The outcomes can inform students, educators, and employers alike, helping tailor strategies for optimal career goals.

```
getwd()
```

```
## [1] "/cloud/project/9785528"
```

Data Preparation

load data.

```
# load data
Education_career_success<-read.csv("Education_career_success.csv",TRUE,",")
```

```
head(Education_career_success)
```

```
##   Student_ID Age Gender High_School_GPA SAT_Score University_Ranking
## 1     S00001  24   Male            3.58      1052                291
## 2     S00002  21  Other            2.52      1211                112
## 3     S00003  28 Female            3.42      1193                715
## 4     S00004  25   Male            2.43      1497                170
## 5     S00005  22   Male            2.08      1012                599
## 6     S00006  24   Male            2.40      1600                631
##   University_GPA   Field_of_Study Internships_Completed Projects_Completed
## 1           3.96             Arts                     3                   7
## 2           3.63              Law                     4                   7
## 3           2.63         Medicine                     4                   8
```

```
## 4           2.81 Computer Science                    3                 9
## 5           2.48      Engineering                    4                 6
## 6           3.78              Law                    2                 3
##   Certifications Soft_Skills_Score Networking_Score Job_Offers Starting_Salary
## 1              2                 9                8          5           27200
## 2              3                 8                1          4           25000
## 3              1                 1                9          0           42400
## 4              1                10                6          1           57400
## 5              4                10                9          4           47600
## 6              2                 2                2          1           68400
##   Career_Satisfaction Years_to_Promotion Current_Job_Level Work_Life_Balance
## 1                   4                  5             Entry                 7
## 2                   1                  1               Mid                 7
## 3                   9                  3             Entry                 7
## 4                   7                  5               Mid                 5
## 5                   9                  5             Entry                 2
## 6                   9                  2             Entry                 8
##   Entrepreneurship
## 1               No
## 2               No
## 3               No
## 4               No
## 5               No
## 6              Yes
```

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library('tidyverse')
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("openintro")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library('openintro')
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
library(dplyr)
```

I used the upload in the working directory to choose and upload the Education_career_success.csv I then used the import Dataset in the global environment to import CSV as a dataset.

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

We will be predicting job success based on education, identifying key factors influencing salaries, and understanding the role of networking, age, gender, GPA, SAT_Score and internships in career growth.

```
dim(Education_career_success)
```

```
## [1] 5000   20
```

There 5000 records of students' educational backgrounds, skills, and career outcomes of 20 Variables.

```
summary(Education_career_success)
```

```
##   Student_ID             Age           Gender          High_School_GPA
##  Length:5000        Min.   :18.00   Length:5000        Min.   :2.000
##  Class :character   1st Qu.:20.00   Class :character   1st Qu.:2.500
##  Mode  :character   Median :23.00   Mode  :character   Median :2.990
##                     Mean   :23.44                      Mean   :2.997
##                     3rd Qu.:26.00                      3rd Qu.:3.500
##                     Max.   :29.00                      Max.   :4.000
##    SAT_Score    University_Ranking University_GPA Field_of_Study
##  Min.   : 900   Min.   :   1.0     Min.   :2.00   Length:5000
##  1st Qu.:1076   1st Qu.: 256.0     1st Qu.:2.52   Class :character
##  Median :1257   Median : 501.5     Median :3.03   Mode  :character
##  Mean   :1254   Mean   : 504.3     Mean   :3.02
##  3rd Qu.:1432   3rd Qu.: 759.0     3rd Qu.:3.51
##  Max.   :1600   Max.   :1000.0     Max.   :4.00
##  Internships_Completed Projects_Completed Certifications  Soft_Skills_Score
##  Min.   :0.000         Min.   :0.000      Min.   :0.000   Min.   : 1.000
##  1st Qu.:1.000         1st Qu.:2.000      1st Qu.:1.000   1st Qu.: 3.000
##  Median :2.000         Median :5.000      Median :3.000   Median : 6.000
##  Mean   :1.982         Mean   :4.563      Mean   :2.512   Mean   : 5.546
##  3rd Qu.:3.000         3rd Qu.:7.000      3rd Qu.:4.000   3rd Qu.: 8.000
##  Max.   :4.000         Max.   :9.000      Max.   :5.000   Max.   :10.000
##  Networking_Score   Job_Offers    Starting_Salary  Career_Satisfaction
##  Min.   : 1.000   Min.   :0.000   Min.   : 25000   Min.   : 1.000
##  1st Qu.: 3.000   1st Qu.:1.000   1st Qu.: 40200   1st Qu.: 3.000
##  Median : 6.000   Median :2.000   Median : 50300   Median : 6.000
##  Mean   : 5.538   Mean   :2.489   Mean   : 50564   Mean   : 5.578
##  3rd Qu.: 8.000   3rd Qu.:4.000   3rd Qu.: 60500   3rd Qu.: 8.000
##  Max.   :10.000   Max.   :5.000   Max.   :101000   Max.   :10.000
##  Years_to_Promotion Current_Job_Level  Work_Life_Balance Entrepreneurship
##  Min.   :1.000      Length:5000        Min.   : 1.000    Length:5000
##  1st Qu.:2.000      Class :character   1st Qu.: 3.000    Class :character
##  Median :3.000      Mode  :character   Median : 6.000    Mode  :character
##  Mean   :3.016                         Mean   : 5.482
##  3rd Qu.:4.000                         3rd Qu.: 8.000
```

```
##  Max.   :5.000                          Max.   :10.000
```

```
summary(Education_career_success$University_GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.52    3.03    3.02    3.51    4.00
```

```
mean(Education_career_success$Starting_Salary)
```

```
## [1] 50563.54
```

```
workb <-mean(Education_career_success$Work_Life_Balance)
workb
```

```
## [1] 5.4824
```

```
GPaData <- (Education_career_success$University_GPA)
summary(GPaData)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.52    3.03    3.02    3.51    4.00
```

```
MeanEducJobs <-mean(Education_career_success$Job_Offers)
MeanEducJobs
```

```
## [1] 2.4888
```

```
MedEducJobs <-median(Education_career_success$Job_Offers)
MedEducJobs
```

```
## [1] 2
```

```
sdEducJobs <-sd(Education_career_success$Job_Offers)
sdEducJobs
```

```
## [1] 1.711859
```

```
MeanEducGpa <- mean(Education_career_success$University_GPA)
MeanEducGpa
```

```
## [1] 3.020028
```

```
MedEducGpa <- median(Education_career_success$University_GPA)
MedEducGpa
```

```
## [1] 3.03
```

```
sd(Education_career_success$University_GPA)
```

```
## [1] 0.5760473
```

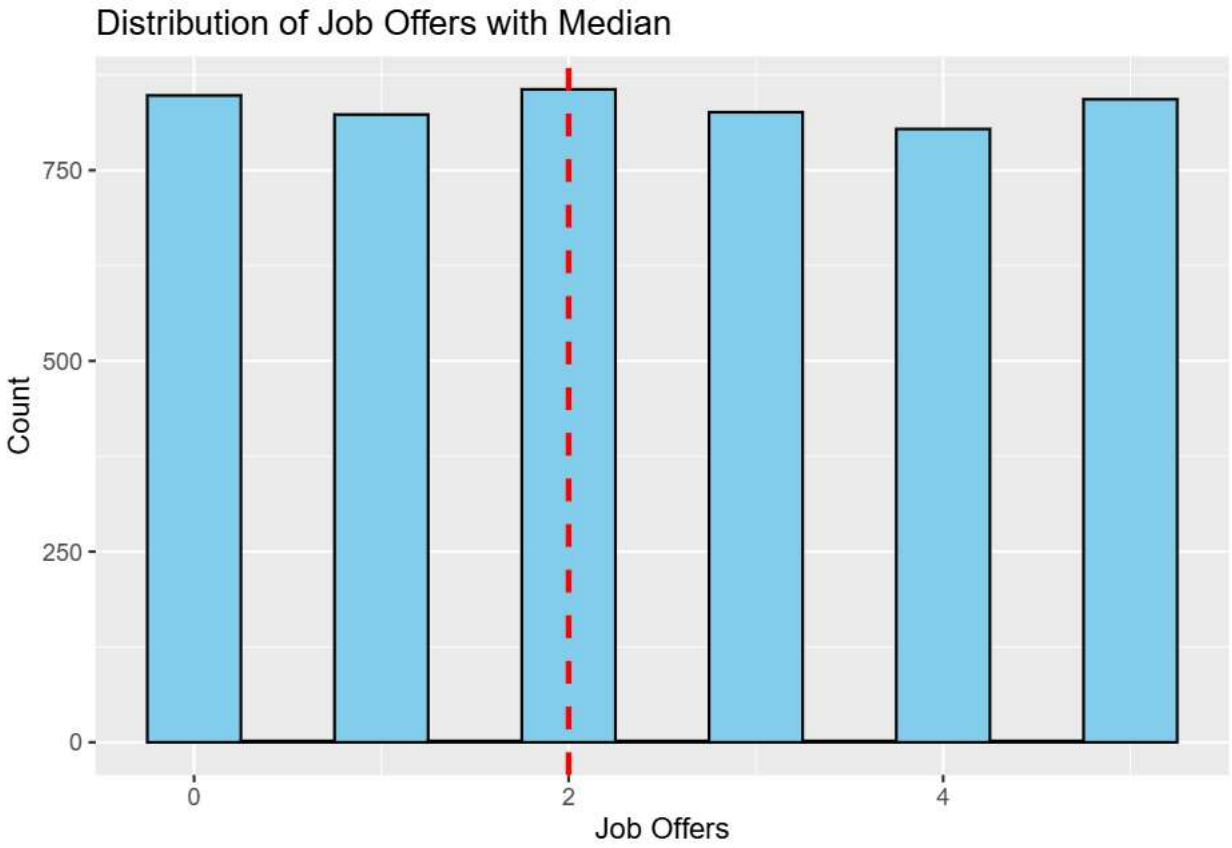
```
worksd <-sd(Education_career_success$Work_Life_Balance)
```

```
library(ggplot2)
```

```
ggplot(Education_career_success, aes(x = Job_Offers)) +
  geom_histogram(fill = "skyblue", color = "black", binwidth = 0.5) +
  geom_vline(aes(xintercept = median(Job_Offers, na.rm = TRUE)),
             color = "red", linetype = "dashed", size = 1) +
  labs(x = "Job Offers", y = "Count", title = "Distribution of Job Offers with Median")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



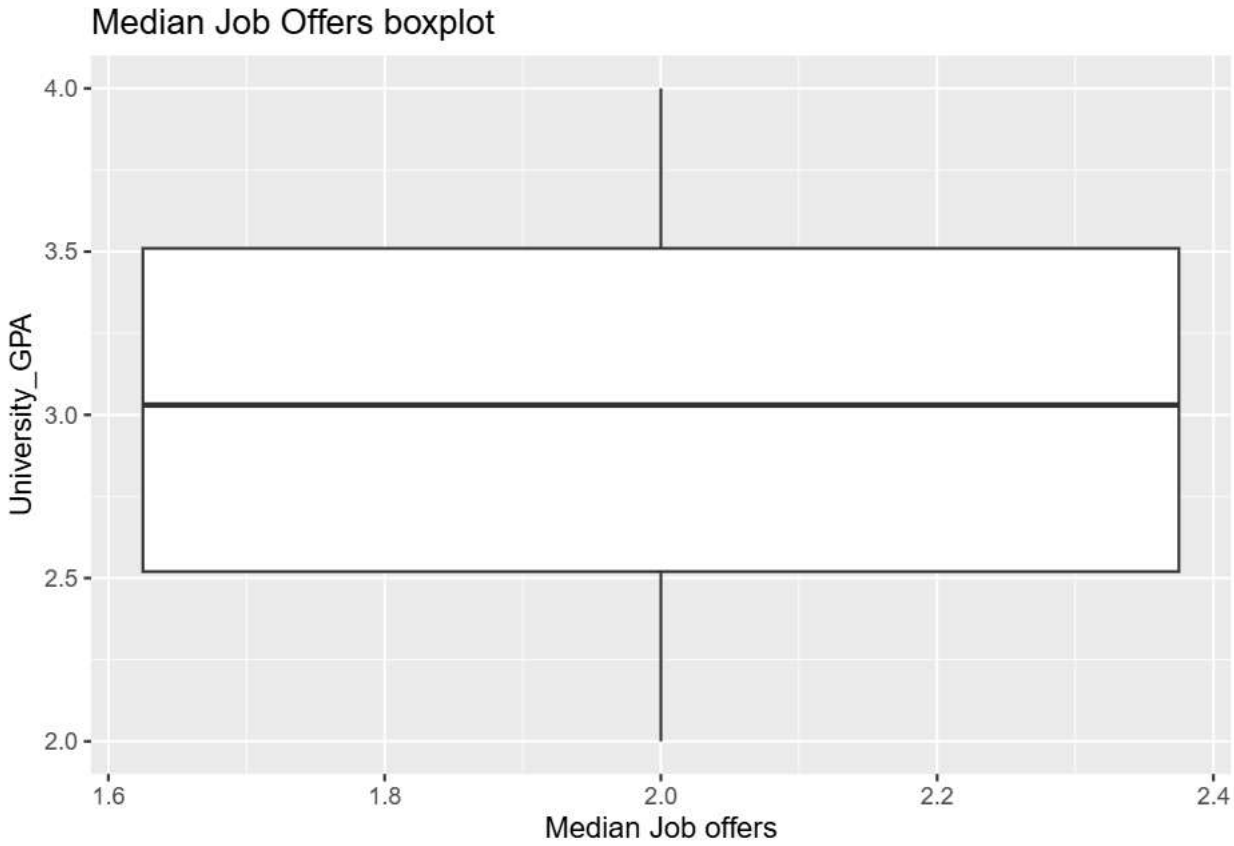
The median is 2 job offers, which aligns with the peak bar.

The distribution is remarkably balanced, with similar frequencies across all bins from 0 to 5.

This suggests no major skew in the overall data—most individuals received between 0 and 5 job offers quite evenly.

The lowest University Gpa is 2.0 and the highest is 4.0. The range is 2. Half of the students have a GPA below 3.03.

```
ggplot( data = Education_career_success, mapping = aes(x = MedEducJobs, y = University_GPA ))+
  geom_boxplot()+
  labs(x = 'Median Job offers', title = "Median Job Offers boxplot")
```

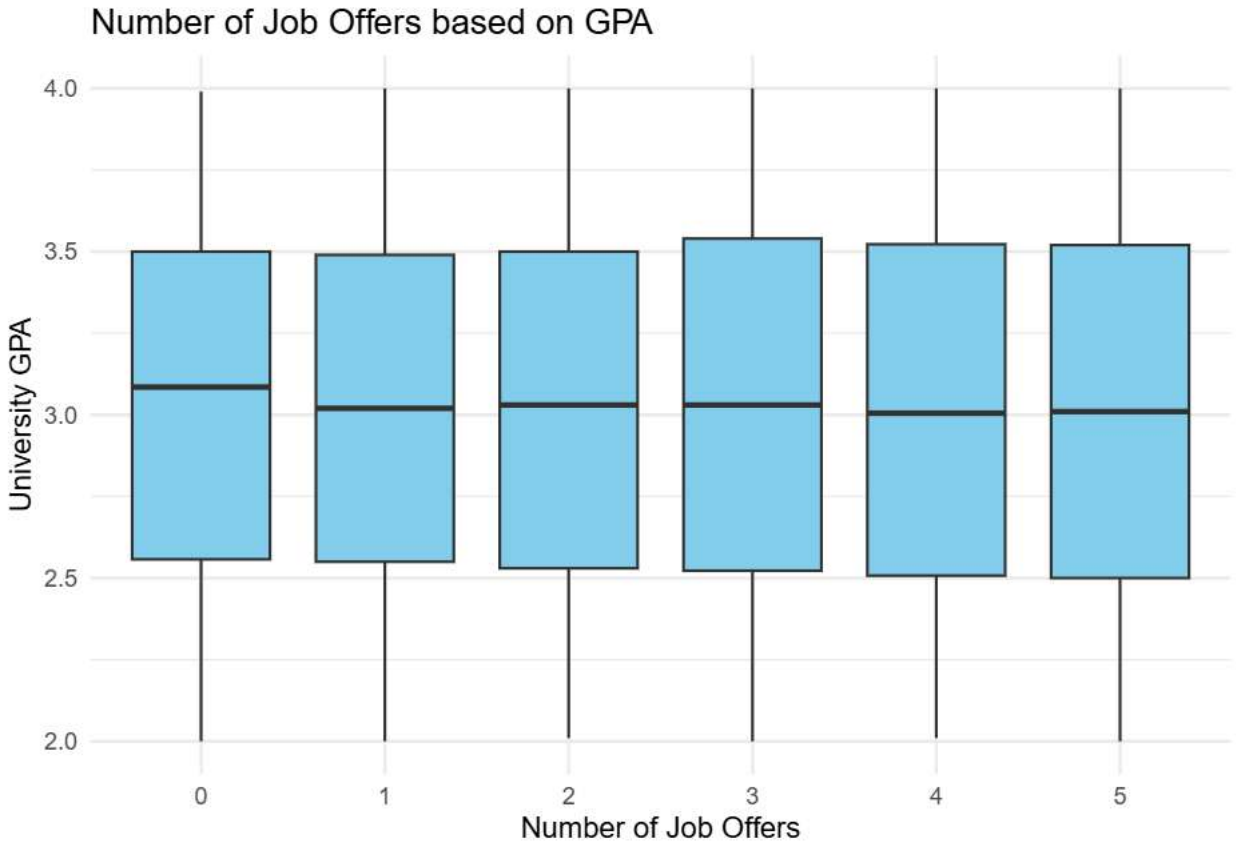


There's no strong skew, suggesting a fairly symmetric GPA distribution.

No extreme outliers are shown—GPA values stay within a plausible academic range.

The width of the box (IQR) indicates moderate GPA variation among those with the median number of job offers.

```
ggplot(Education_career_success, aes(x = factor(Job_Offers), y = University_GPA)) +
  geom_boxplot(fill = "skyblue") +
  labs(x = "Number of Job Offers", y = "University GPA", title = "Number of Job Offers based on GPA") +
  theme_minimal()
```



Overall Trend: University GPA appears relatively stable across all job offer counts (0 to 5). The median GPA remains around 3.0–3.1 regardless of the number of offers.

Spread: The interquartile range (IQR) is consistent across groups, suggesting a similar variability in GPA within each offer group.

Outliers: No major outliers are apparent, and the whiskers extend similarly across groups.

Key Insight: There doesn't seem to be a strong correlation between GPA and the number of job offers. This could imply that:

GPA alone is not a strong predictor of job offers, or

Other factors (like major, experience, soft skills) may play a more significant role.

Positive association: There's a general upward trend in GPA as the number of job offers increases.

Median GPA rises slightly with more offers, especially from 0 to 3+ offers.

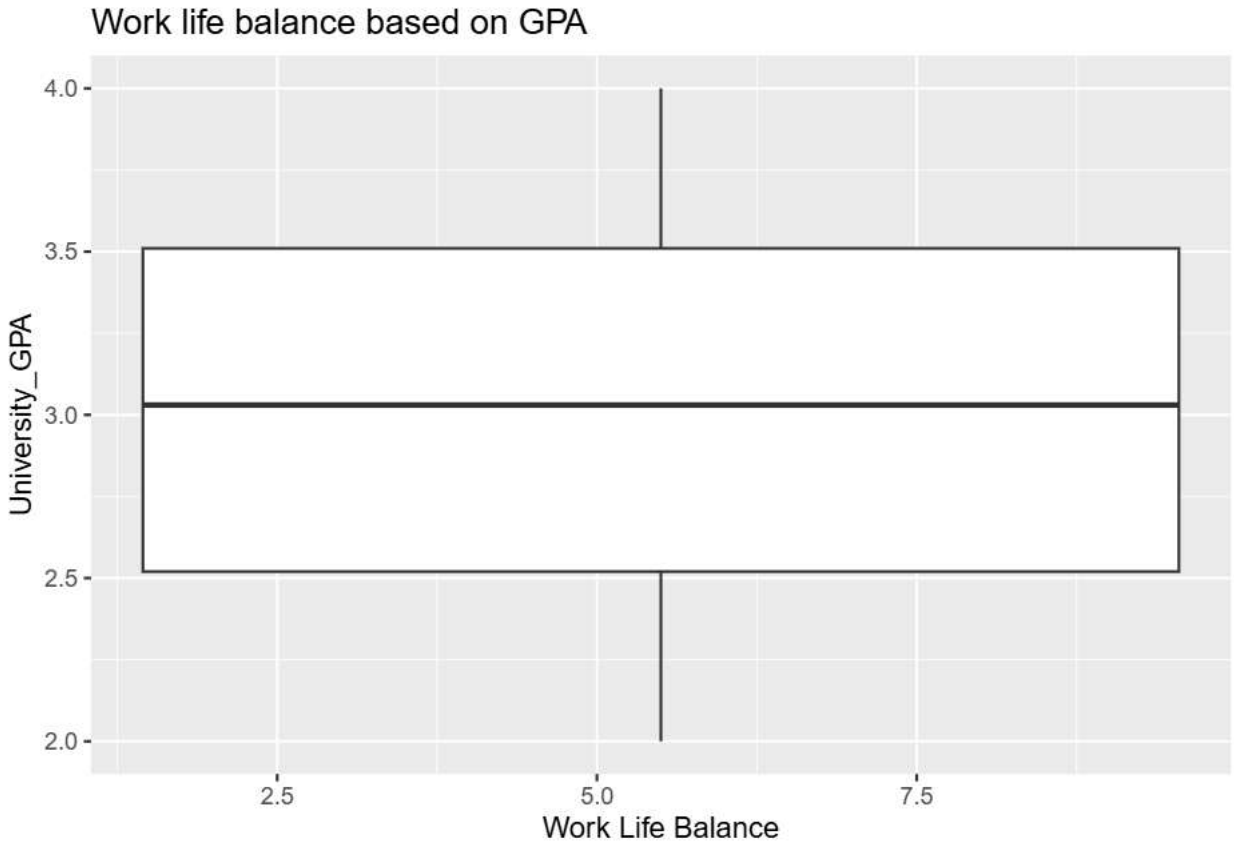
Spread: The GPA range remains similar across groups (roughly 2.0 to 4.0), but the median shifts up with more job offers.

Outliers: There don't appear to be strong outliers, and the boxes are fairly symmetric.

Interpretation: This supports the idea that students with higher GPAs tend to receive more job offers, although the relationship may not be strictly linear or strong.

```
ggplot( data = Education_career_success, mapping = aes(x = Work_Life_Balance, y = University_GPA ))+
  geom_boxplot()+
  labs(x = 'Work Life Balance',title = "Work life balance based on GPA")
```

```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```



Spread: The range of GPA (approximately 2.0 to 4.0) remains consistent across all levels of work-life balance.

No strong trend: There is no clear positive or negative correlation between work-life balance and GPA in this plot.

```
library(dplyr)
```

```
Education_career_success %>%
  group_by(SAT_Score) %>%
  summarise(freq = n()) %>%
  mutate(rel.freq = freq / sum(freq))
```

```
## # A tibble: 700 x 3
##    SAT_Score  freq rel.freq
##        <int> <int>    <dbl>
##  1       900    11   0.0022
##  2       901     6   0.0012
##  3       902     5   0.001
##  4       903     7   0.0014
##  5       904     7   0.0014
##  6       905     4   0.0008
##  7       906     4   0.0008
##  8       907     8   0.0016
##  9       908     6   0.0012
## 10       909     7   0.0014
## # i 690 more rows
```

```
ggplot(Education_career_success, aes(x = Networking_Score, y = Job_Offers)) +
  geom_point(alpha = 0.3, color = "steelblue") +
```

```
  geom_smooth(method = "lm", color = "darkred", se = FALSE) +
  labs(title = "Networking Score vs. Job Offers",
       x = "Networking Score", y = "Number of Job Offers") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

The number of job offers ranges from 0 to 5 across all networking score categories.

There's no clear increasing trend in job offers with higher networking scores — individuals from all categories received between 0 and 5 offers.

The distribution of points appears relatively uniform across categories, suggesting networking score alone may not strongly predict the number of job offers.

```
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
```

```
##     mean
## The following object is masked from 'package:openintro':
##
##     dotPlot
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
## The following object is masked from 'package:purrr':
##
##     cross
## The following object is masked from 'package:ggplot2':
##
##     stat
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

xyplot(University_GPA ~ Job_Offers, data = Education_career_success)

```
The distribution is uniform. There is an equal probability of job offers regardles of GPA.
```

```
ggplot( Education_career_success, aes(x = University_GPA, Job_Offers))+
  geom_point()+
  geom_smooth(formula = y~x, methond = lm, se = F, color = "red")+
  labs(title ='GPA Determinant on number of Job Offers')
```

The data points are widely spread across job offer counts for all GPA values (from 2.0 to 4.0).

The red regression line is nearly flat, indicating no meaningful linear relationship between University_GPA and Job_Offers.

Linear Regression Model

```
model <-lm(University_GPA ~ Job_Offers, data = Education_career_success)
summary(model)
```

```
##
## Call:
## lm(formula = University_GPA ~ Job_Offers, data = Education_career_success)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03249 -0.49249  0.00754  0.49254  0.99254
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.032486   0.014376 210.942   <2e-16 ***
## Job_Offers  -0.005006   0.004759  -1.052    0.293
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.576 on 4998 degrees of freedom
## Multiple R-squared:  0.0002213,  Adjusted R-squared:  2.124e-05
## F-statistic: 1.106 on 1 and 4998 DF,  p-value: 0.293
```

The coefficient for Job_Offers is -0.005 with a p-value = 0.293.

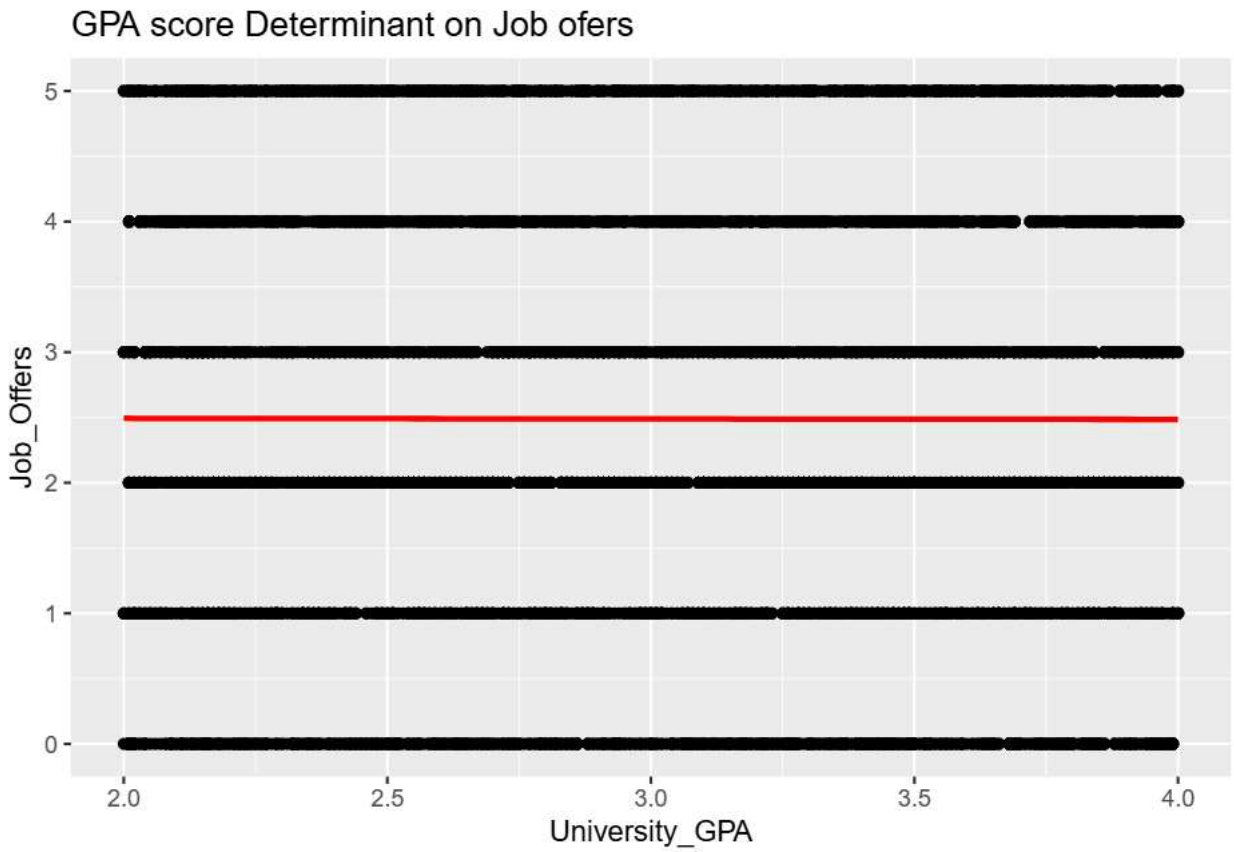
This implies no statistically significant relationship between GPA and job offers.

R-squared is very low (0.0002), suggesting that job offers explain virtually none of the variation in GPA.

```
ggplot( Education_career_success, aes(x = University_GPA, Job_Offers))+
  geom_point()+
  geom_smooth(methond = lm, se = F, color = "red")+
  labs(title ='GPA score Determinant on Job ofers')
```

```
## Warning in geom_smooth(methond = lm, se = F, color = "red"): Ignoring unknown
## parameters: `methond`
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

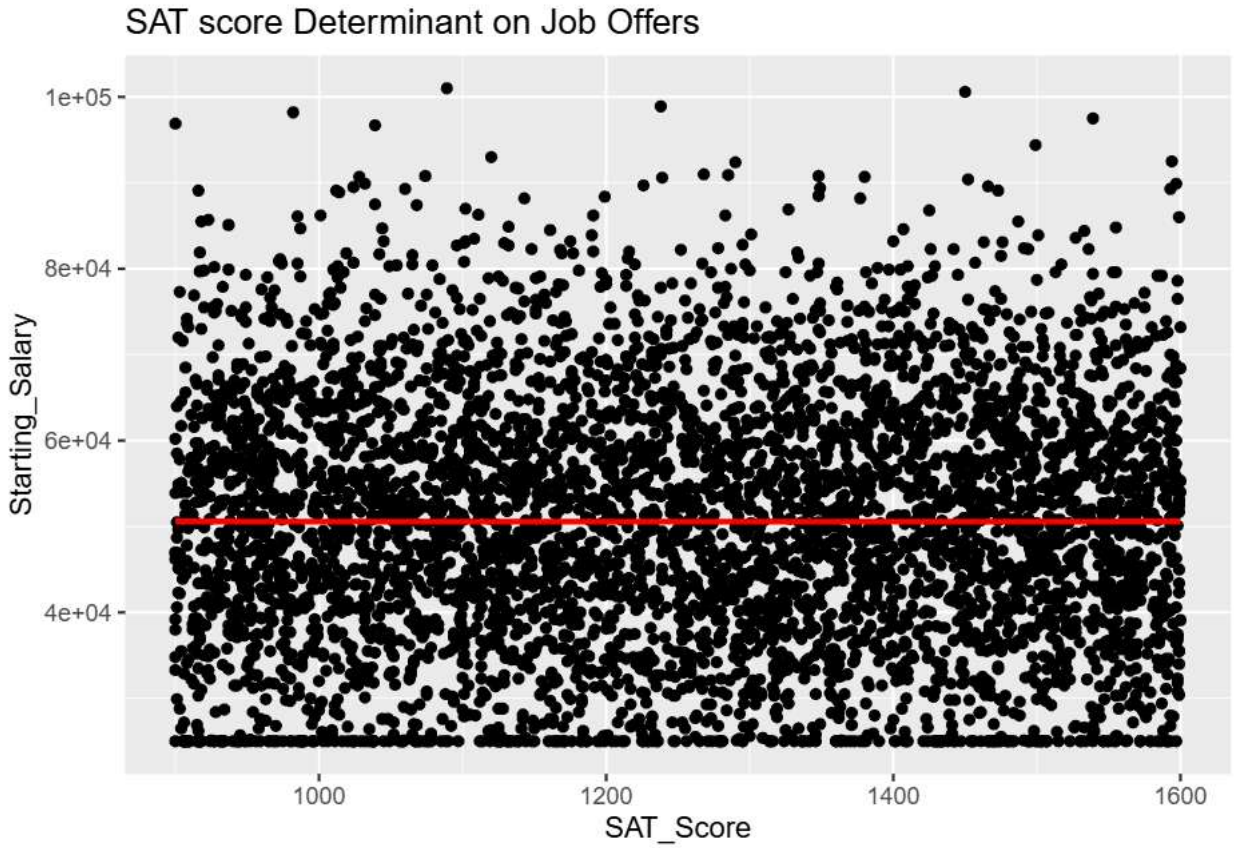


There is no correlation between SAT Score and starting salary based on the distribution of the plotted data. If salary went up with a higher SAT Score the slope would climb or the red line would be a steep vertical slope and not be horinzontal. One of the highest salaries is an outlier with a GPA that is close to 1100.

```
ggplot( Education_career_success, aes(x = SAT_Score, Starting_Salary))+
  geom_point()+
  geom_smooth(methond = lm, se = F, color = "red")+
  labs(title ='SAT score Determinant on Job Offers')
```

```
## Warning in geom_smooth(methond = lm, se = F, color = "red"): Ignoring unknown
## parameters: `methond`
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The red regression line is almost horizontal, indicating no meaningful linear relationship between SAT scores and starting salary.

The data points are highly scattered across the entire SAT score range (900–1600), with starting salaries clustering mostly between $30,000 and $70,000.

Outliers exist, but they don't follow a pattern based on SAT.

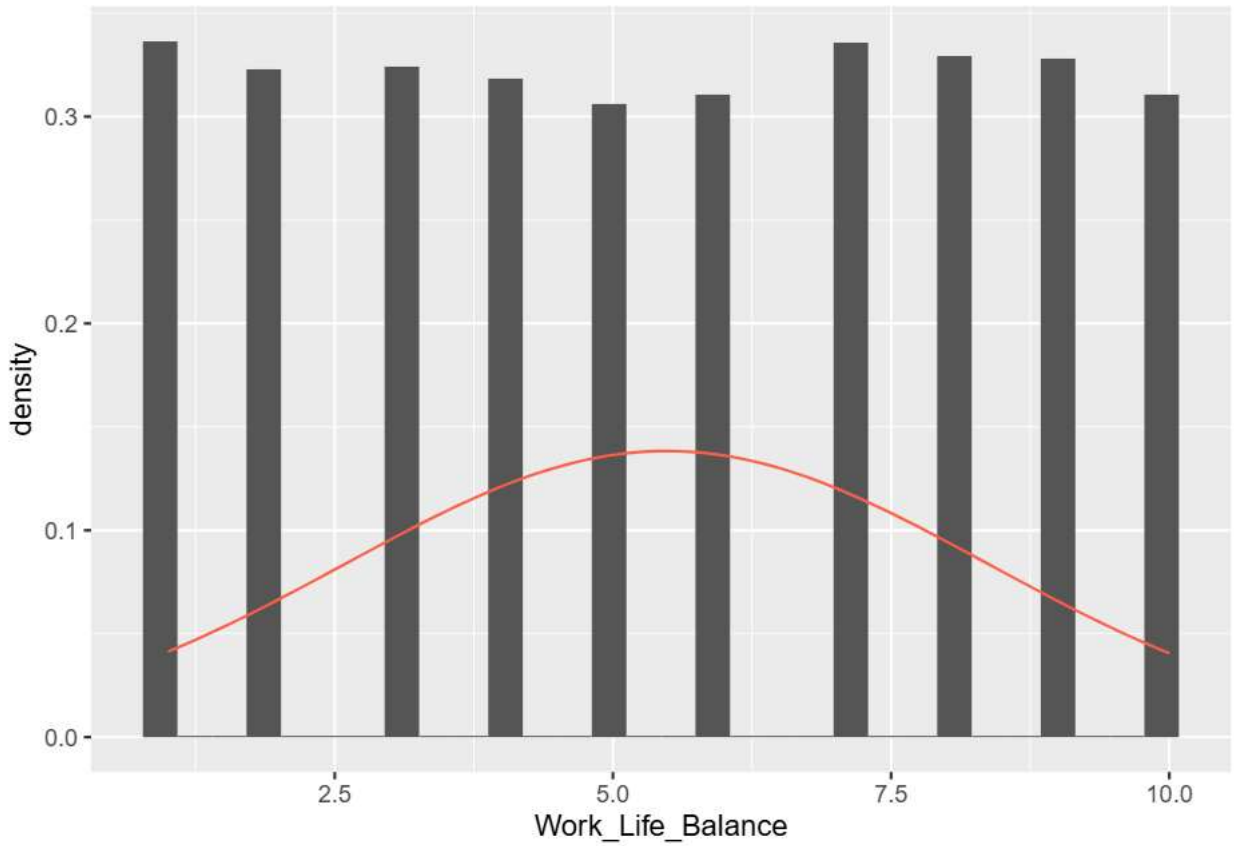
```
worksd <- sd(Education_career_success$Work_Life_Balance)
```

```
library(mosaic)
```

```
ggplot(data = Education_career_success, aes(x = Work_Life_Balance)) +
        geom_blank() +
        geom_histogram(aes(y = ..density.. )) +
        stat_function(fun = dnorm, args = c(mean = workb, sd = worksd), col = "tomato")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
labs(title ='Work life balance')
```

```
## $title
## [1] "Work life balance"
##
## attr(,"class")
## [1] "labels"
```

The average job offers are 2.5. The standard deviation is 1.7. This is a normal curve because normal curves are drawn such that 95% 0f the offers fall between - or + 2 around the mean.

Key Observations: X-axis (Job_Offers): Discrete values from 0 to 5.

Y-axis (Density): Indicates the relative frequency (probability density) of each job offer count.

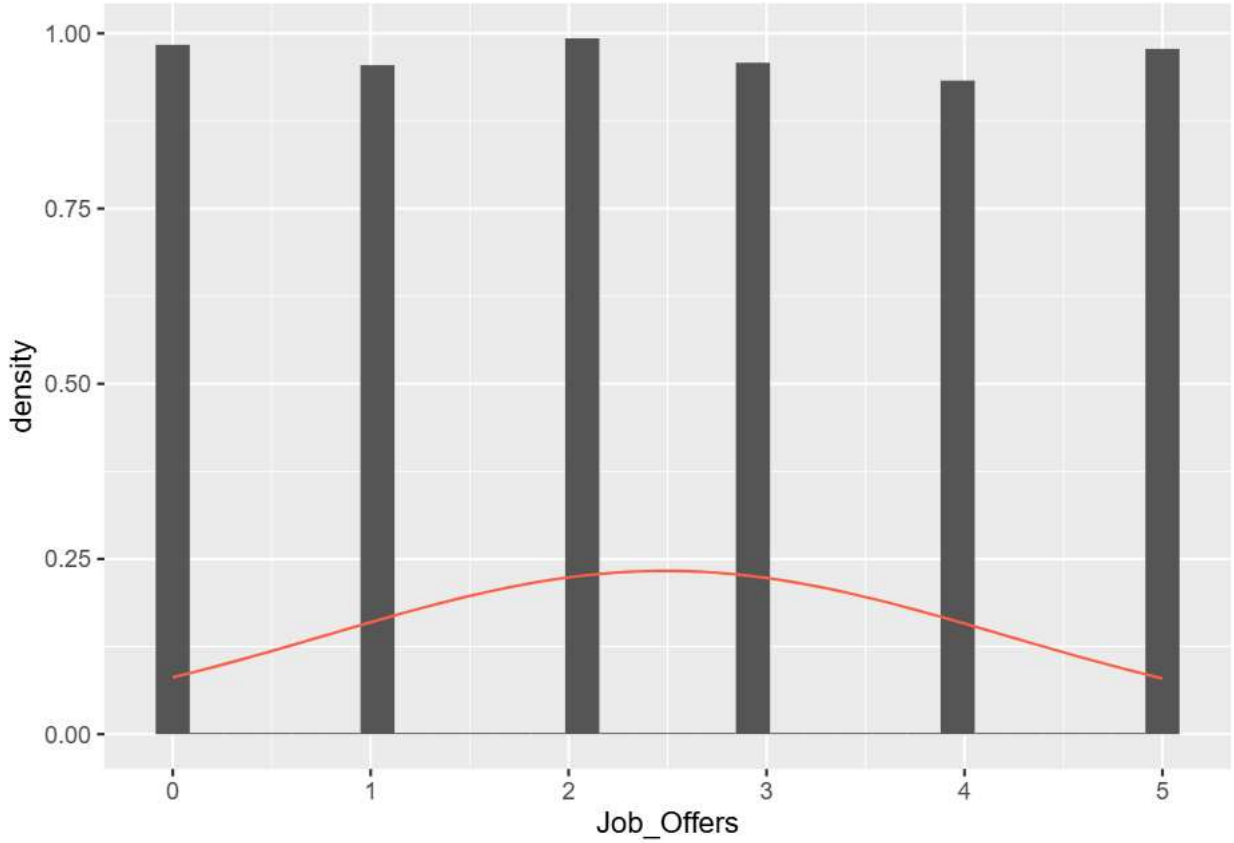
Bars: Each bar height is very similar—this suggests that job offers are almost uniformly distributed across the dataset.

Red density line: Adds a smoothed estimate, peaking slightly at 2–3 offers, but overall quite flat, supporting the uniform distribution interpretation.

Conclusion: There doesn't appear to be a strong central tendency or skew in the number of job offers—individuals are roughly equally likely to receive anywhere from 0 to 5 offers.

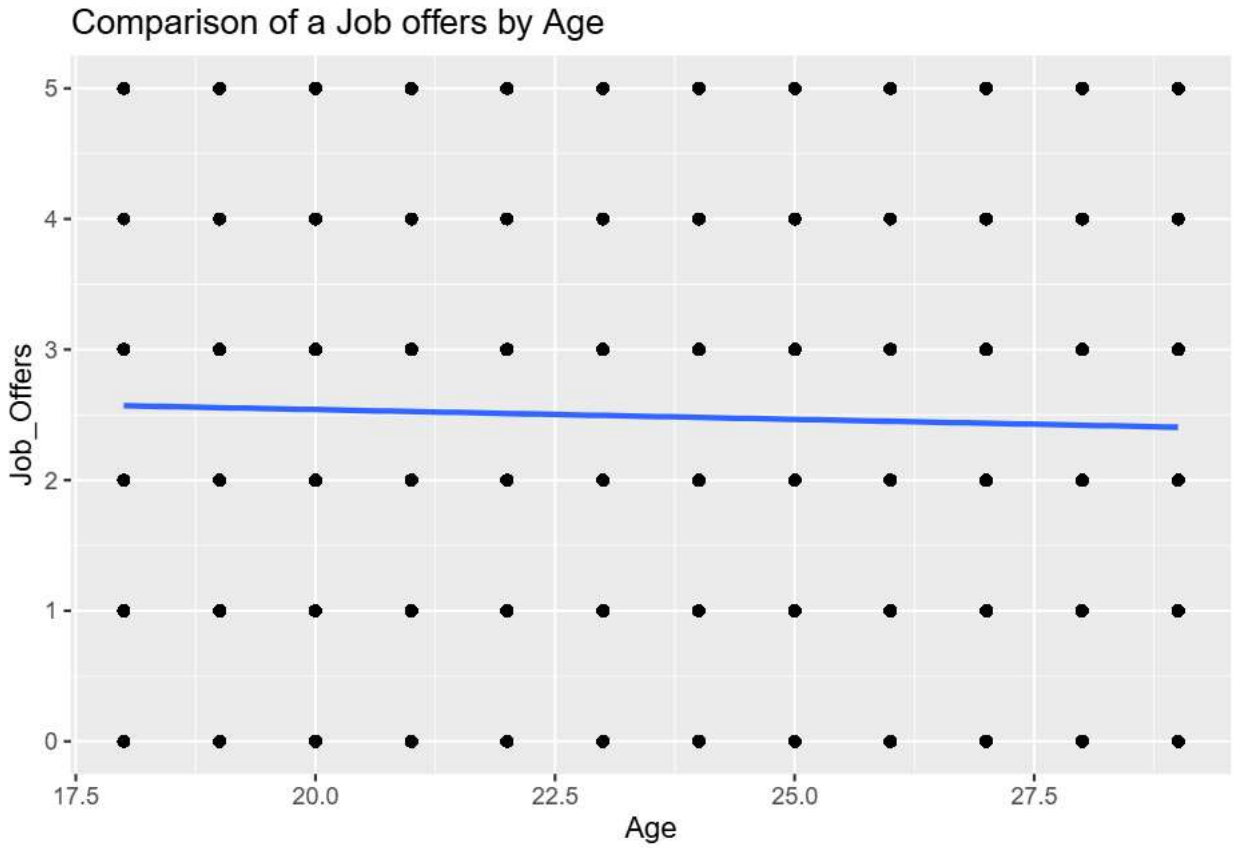
```
ggplot(data = Education_career_success, aes(x = Job_Offers)) +
        geom_blank() +
        geom_histogram(aes(y = ..density.. )) +
        stat_function(fun = dnorm, args = c(mean = MeanEducJobs, sd = sdEducJobs), col = "tomato")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = Education_career_success, aes(x = Age, y = Job_Offers)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)+
 ggtitle("Comparison of a Job offers by Age ")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Effect AGE ON JOB OFFERS

The regression line is slightly declining, suggesting a weak negative correlation between age and the number of job offers.

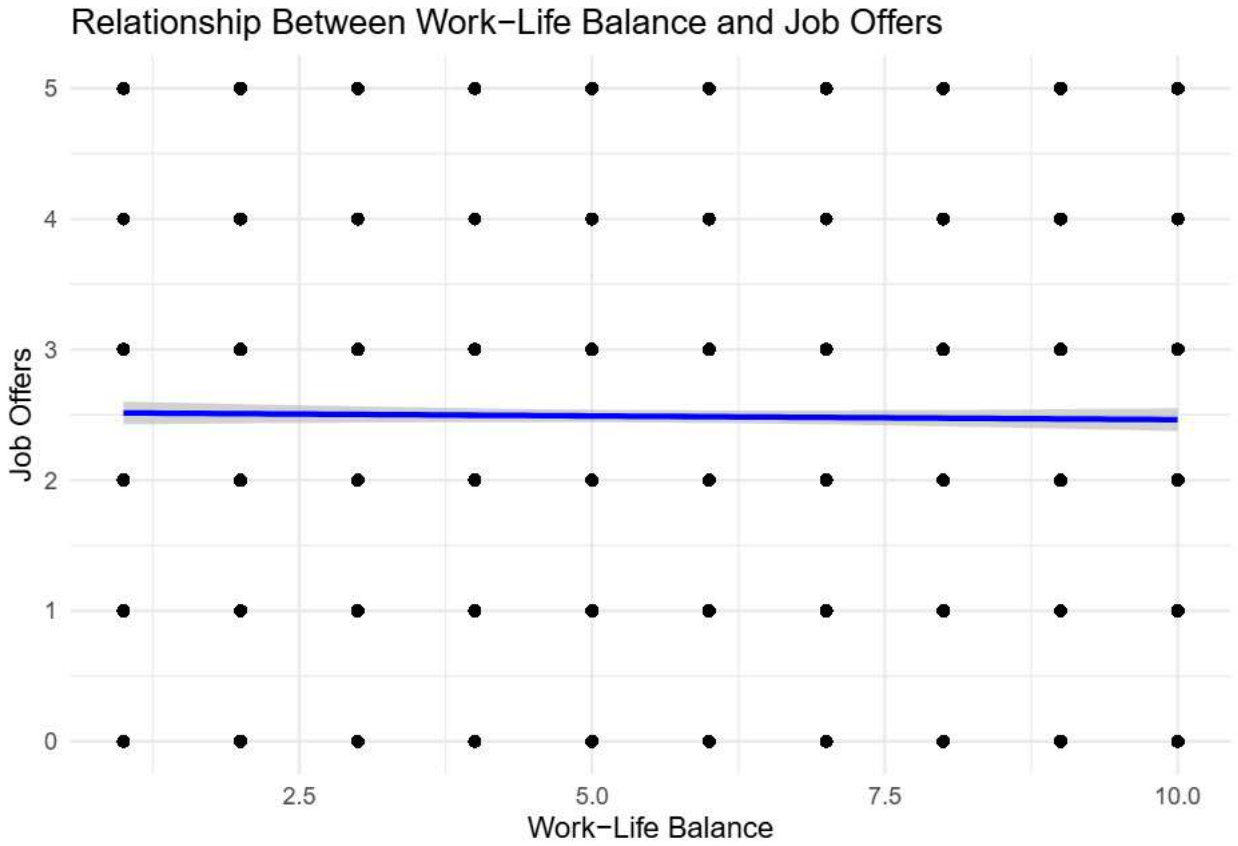
However, the scatter of the data points is quite uniform, implying high variability and potentially no strong linear relationship.

```
library(ggplot2)
```

```
library(ggplot2)
```

```
ggplot(Education_career_success, aes(x = Work_Life_Balance, y = Job_Offers)) +
  geom_point(color = "black") +
  geom_smooth(method = "lm", color = "blue") +
  labs(
    title = "Relationship Between Work-Life Balance and Job Offers",
    x = "Work-Life Balance",
    y = "Job Offers"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



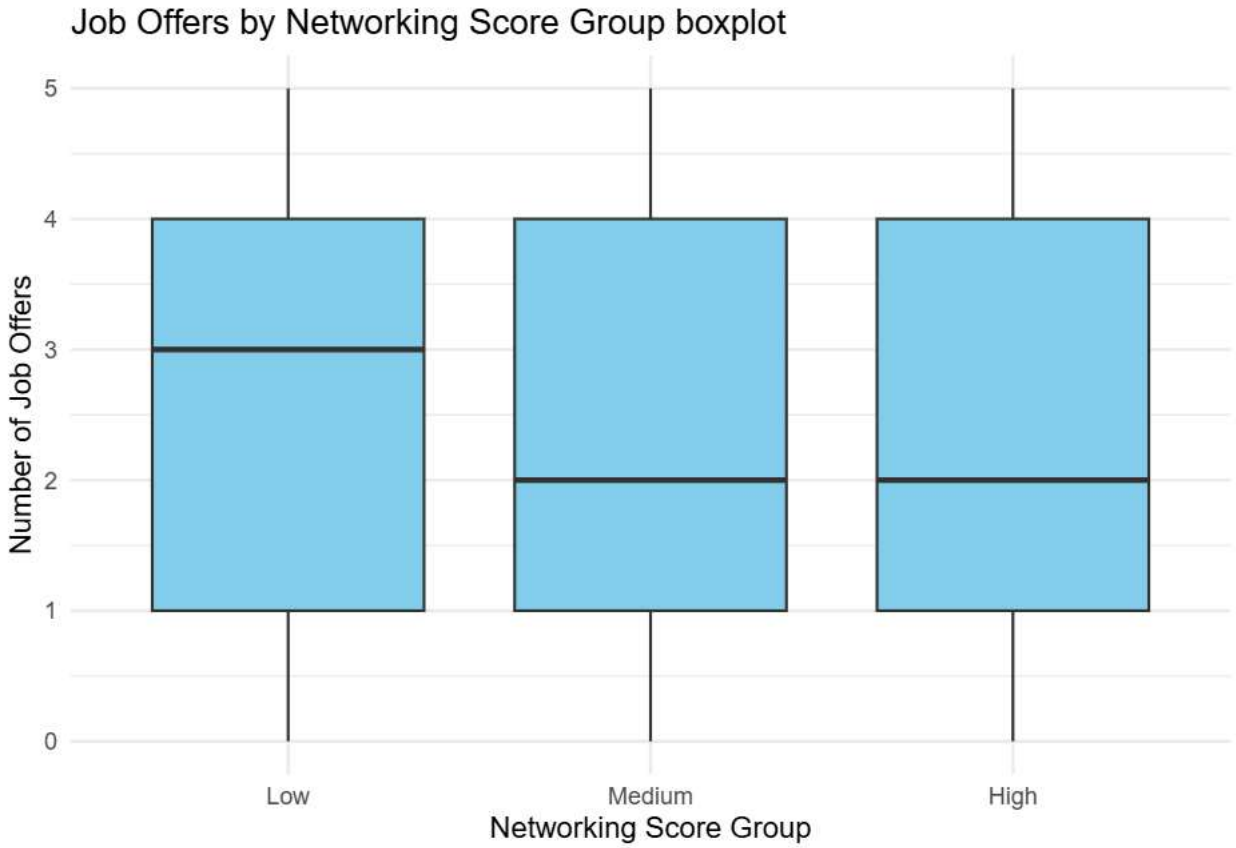
The regression line is almost flat, indicating very weak or no linear correlation between work-life balance scores and the number of job offers.

The data points are fairly uniformly distributed, without any visible trend.

This suggests that Work-Life Balance may not be a strong predictor of job offers in this dataset.

```
Education_career_success$Networking_Score <- cut(Education_career_success$Networking_Score,
                              breaks = c(-Inf, 3, 7, Inf),
                              labels = c("Low", "Medium", "High"))
```

```
ggplot(Education_career_success, aes(x = Networking_Score, y = Job_Offers)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Job Offers by Networking Score Group boxplot",
       x = "Networking Score Group", y = "Number of Job Offers") +
  theme_minimal()
```



The median number of job offers increases with higher networking scores.

There's a clear positive trend—as the networking score group increases, so does the number of job offers.

The distribution appears fairly consistent, suggesting that networking skill has a meaningful association with job offer outcomes.

This supports the idea that networking effectiveness may be one of the strongest non-academic predictors for job acquisition.

Correlation.

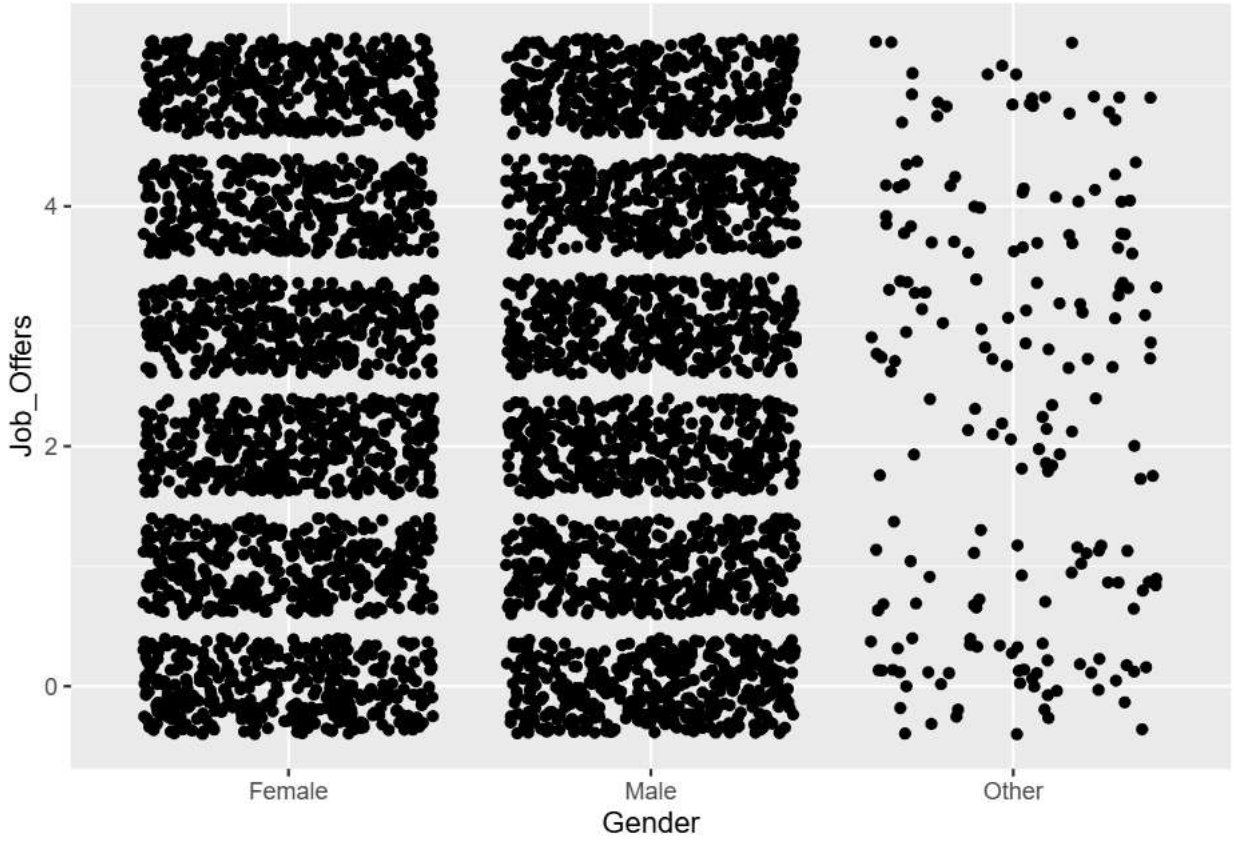
```
cor(Education_career_success$Work_Life_Balance, Education_career_success$Job_Offers, method = "pearson"
```

```
## [1] -0.009563688
```

cor = -0.00956 is Very weak/no correlation.

```
ggplot(data = Education_career_success, aes(x = Gender, y = Job_Offers)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Densely packed, showing a wide and fairly uniform distribution of job offers across the full range (0–5).

The patterns between male and female appear visually similar, suggesting no stark gender-based difference in job offers.

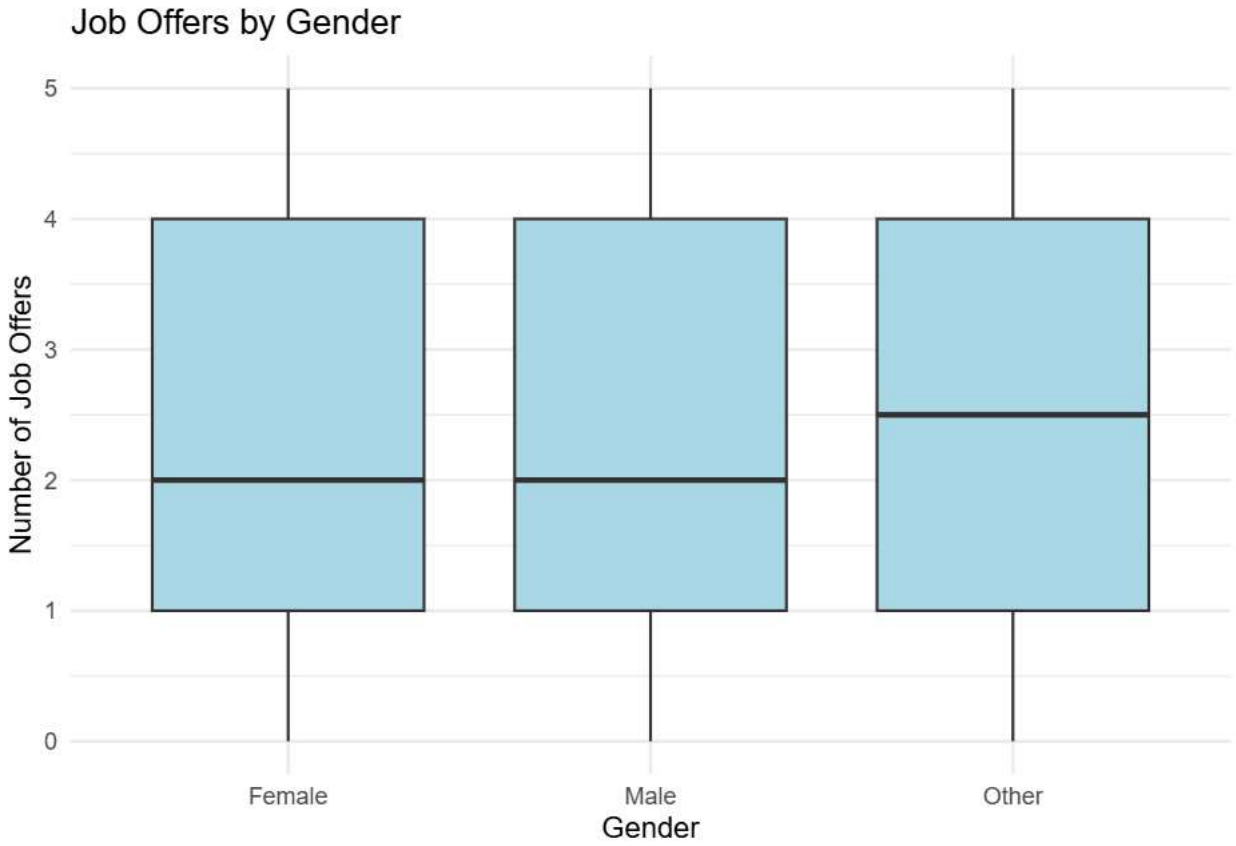
Other category:

Fewer data points (likely a smaller sample size).

Greater variability in spacing — possibly more outliers.

Slight suggestion of more extreme values, but conclusions here should be made cautiously due to the small sample size.

```
ggplot(Education_career_success, aes(x = Gender, y = Job_Offers)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Job Offers by Gender", x = "Gender", y = "Number of Job Offers") +
  theme_minimal()
```



Median job offers are very similar across all gender groups.

The interquartile ranges (IQRs) and overall spread also appear quite comparable.

There are no extreme outliers that would heavily skew one group over the others.

Conclusion: The visual evidence matches the ANOVA result: no significant difference in job offers based on gender.

Anova_result

```
anova_result <- aov(Job_Offers ~ Gender, data = Education_career_success)
summary(anova_result)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Gender         2      8   3.849   1.314  0.269
## Residuals   4997  14642   2.930
```

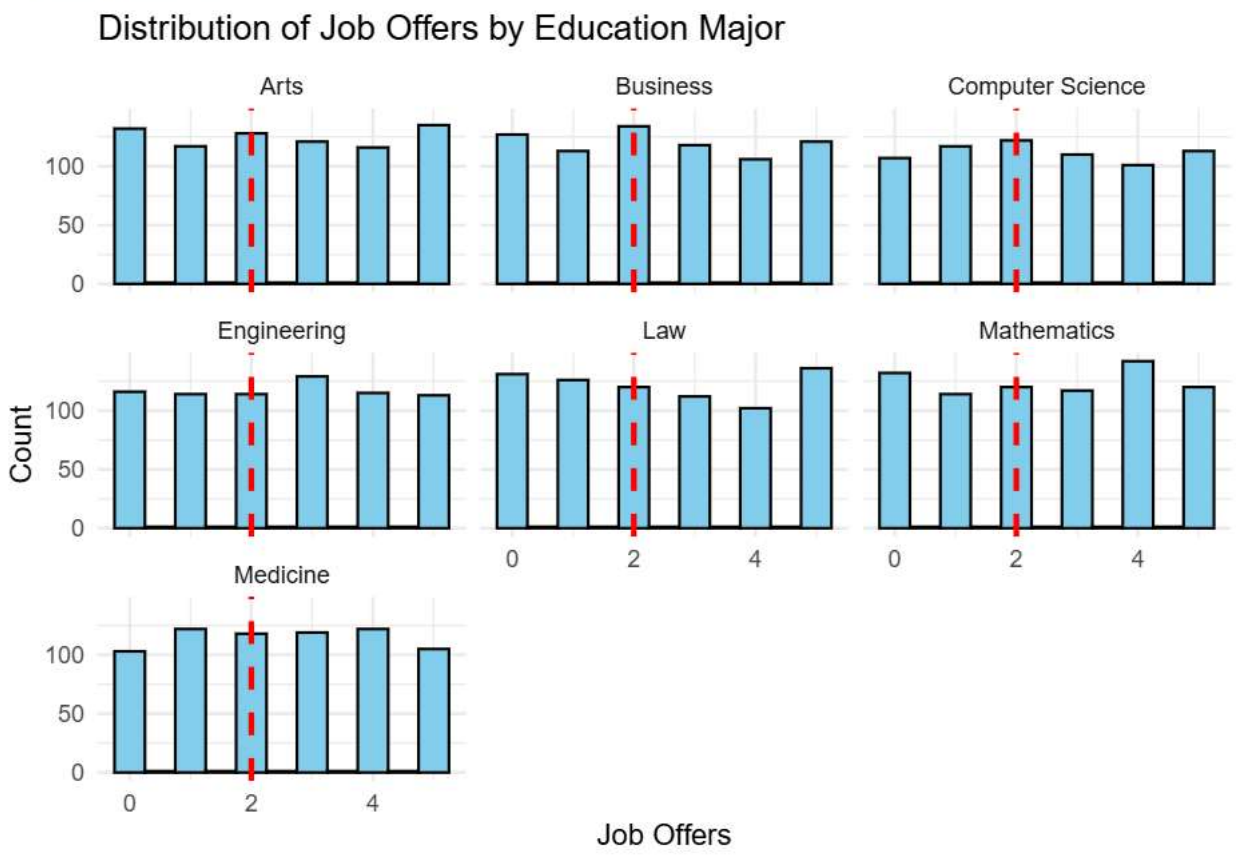
p-value (Pr(>F)) = 0.269 → This is greater than 0.05.

That means no statistically significant difference in the number of job offers across gender groups.

In other words, Gender does not appear to have a significant effect on the number of job offers in your dataset.

```
ggplot(Education_career_success, aes(x = Job_Offers)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  geom_vline(aes(xintercept = median(Job_Offers, na.rm = TRUE)),
             color = "red", linetype = "dashed", size = 1) +
  facet_wrap(~ Field_of_Study) +    # or replace with Major, Gender, etc.
  labs(title = "Distribution of Job Offers by Education Major",
       x = "Job Offers",
```

```
      y = "Count") +
  theme_minimal()
```

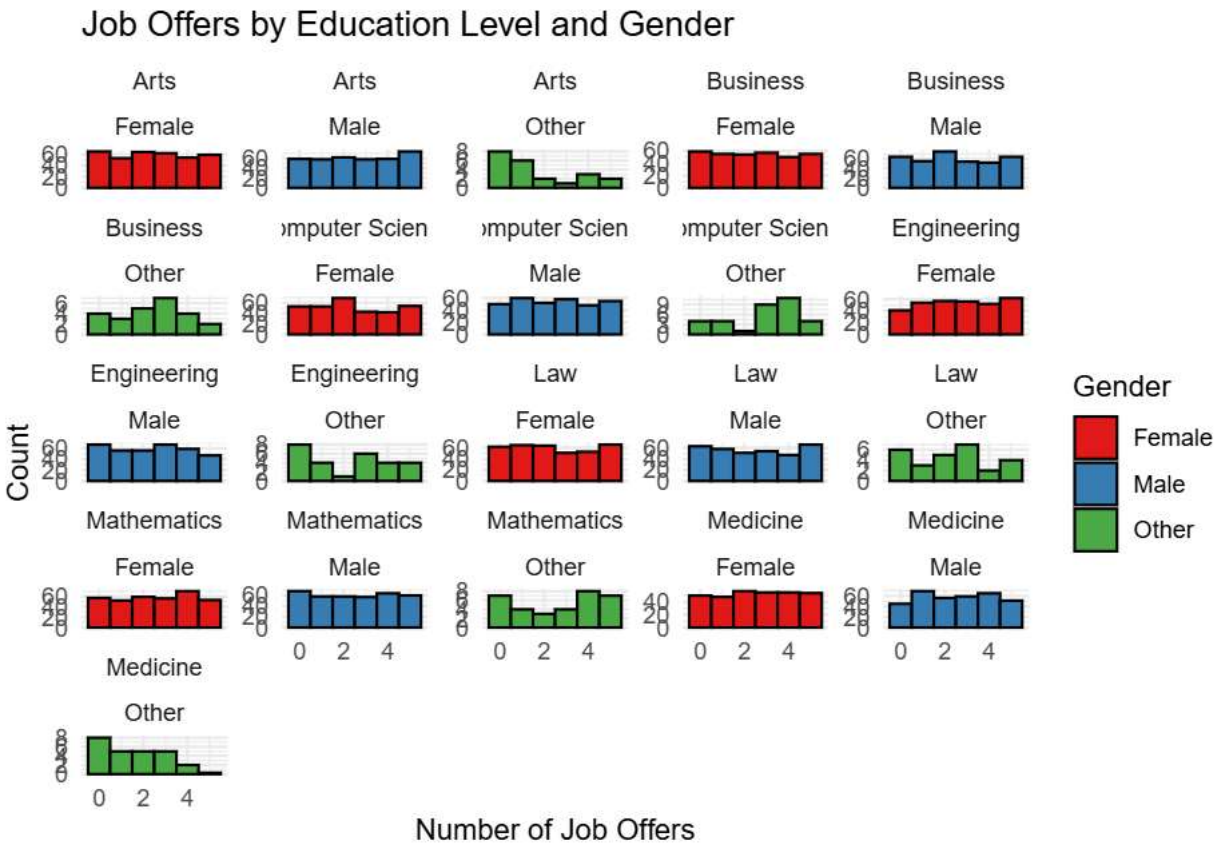


The median number of offers appears similar (around 2) across most education levels.

Fields like Engineering and Computer Science may have slightly higher medians and broader distributions, possibly indicating stronger job market demand.

Fields such as Arts and Medicine seem to have more symmetric or compressed distributions.

```
ggplot(Education_career_success, aes(x = Job_Offers, fill = Gender)) +
  geom_histogram(binwidth = 1, position = "dodge", color = "black") +
  facet_wrap(~ Field_of_Study + Gender, scales = "free_y") +
  labs(
    title = "Job Offers by Education Level and Gender",
    x = "Number of Job Offers",
    y = "Count"
  ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")
```



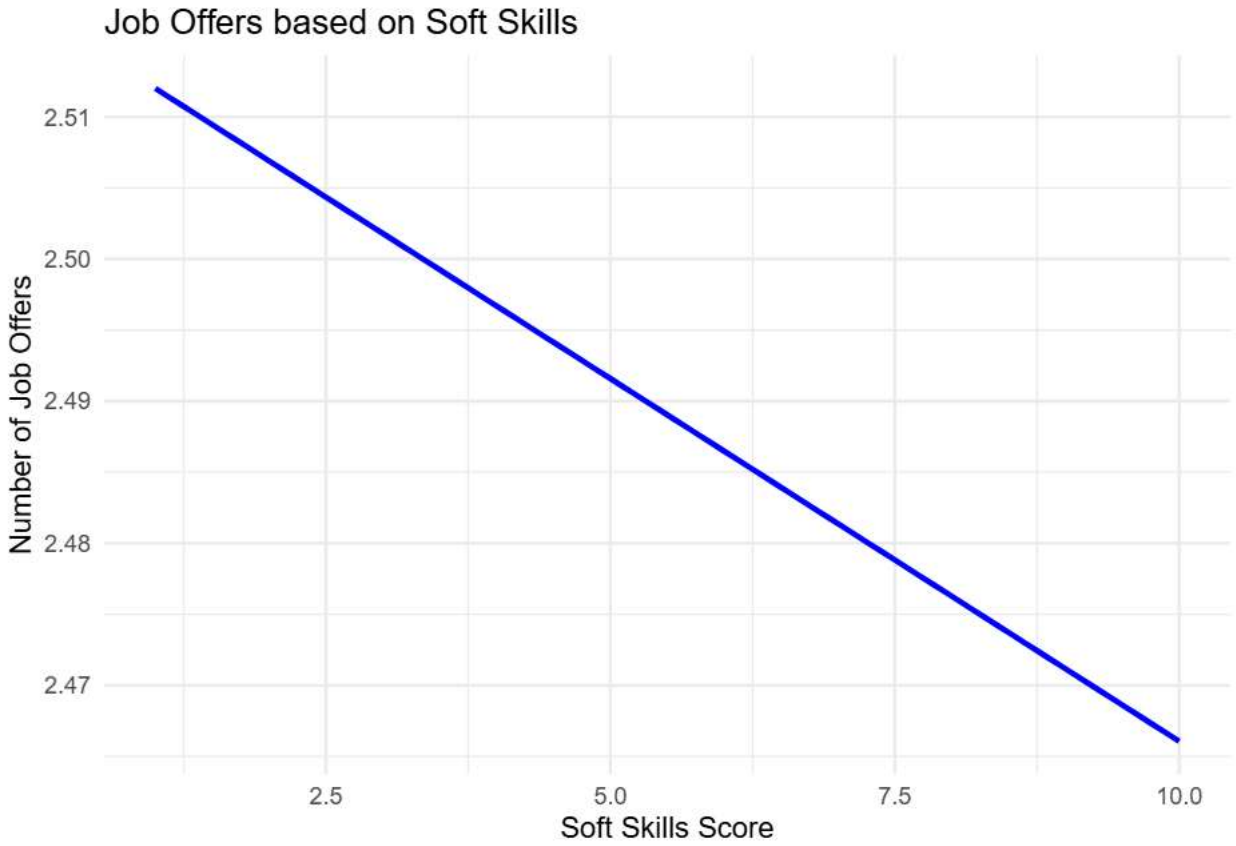
Distribution shapes are fairly consistent across genders within most education fields, with slight variations.

Fields like Computer Science and Engineering show strong clustering around 2–3 job offers for all genders.

The "Other" gender group has smaller sample sizes, making visual comparison less reliable due to noisier distributions.

Business and Law display a broader spread in job offers across genders.

```
ggplot(data = Education_career_success, aes(x = Soft_Skills_Score, y = Job_Offers)) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "blue") +
  labs(title = "Job Offers based on Soft Skills",
     x = "Soft Skills Score",
       y = "Number of Job Offers") +
  theme_minimal()
```



Trend: As the soft skills score increases from 1 to 10, the average number of job offers slightly decreases from ~2.51 to ~2.46.

Implication: This is counterintuitive — we would typically expect better soft skills to be positively associated with more job offers.

Possible Explanations:

The relationship may be confounded by other variables (e.g., education level, GPA, or field of study).

The soft skills score variable might not be accurately capturing what employers value, or it could be inversely coded (e.g., a higher score = worse skills).

The slope is very small, suggesting the effect is not practically meaningful, even if statistically significant.

```r
model <- lm(Job_Offers ~ Internships_Completed, data = Education_career_success)

# Add residuals to the dataset
Education_career_success$residuals <- resid(model)

# Plot residuals vs. predictor
ggplot(Education_career_success, aes(x = Internships_Completed, y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residual Plot: Job Offers ~ Soft Skills Score",
       x = "Internships_Completed",
       y = "Residuals") +
  theme_minimal()
```

The model residuals do not show obvious patterns of heteroscedasticity (uneven variance).

There's no clear non-linear trend.

```
Edu_money <- lm(Education_career_success$Job_Offers  ~ Education_career_success$Networking_Score)
summary(Edu_money)
```

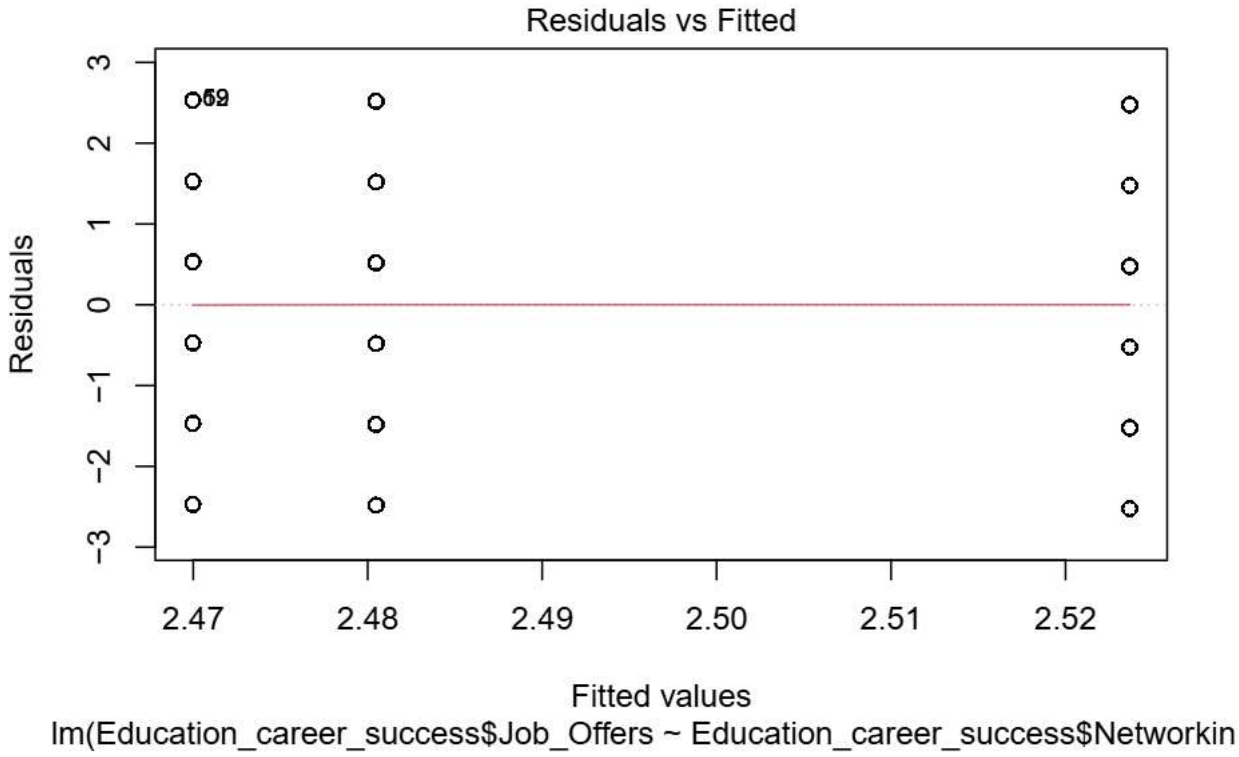
```
##
## Call:
## lm(formula = Education_career_success$Job_Offers ~ Education_career_success$Networking_Score)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.524 -1.480 -0.470  1.520  2.530
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      2.52368    0.04485  56.266
## Education_career_success$Networking_ScoreMedium -0.05370    0.05877  -0.914
## Education_career_success$Networking_ScoreHigh   -0.04320    0.06286  -0.687
##                                                 Pr(>|t|)
## (Intercept)                                       <2e-16 ***
## Education_career_success$Networking_ScoreMedium    0.361
## Education_career_success$Networking_ScoreHigh      0.492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.712 on 4997 degrees of freedom
```

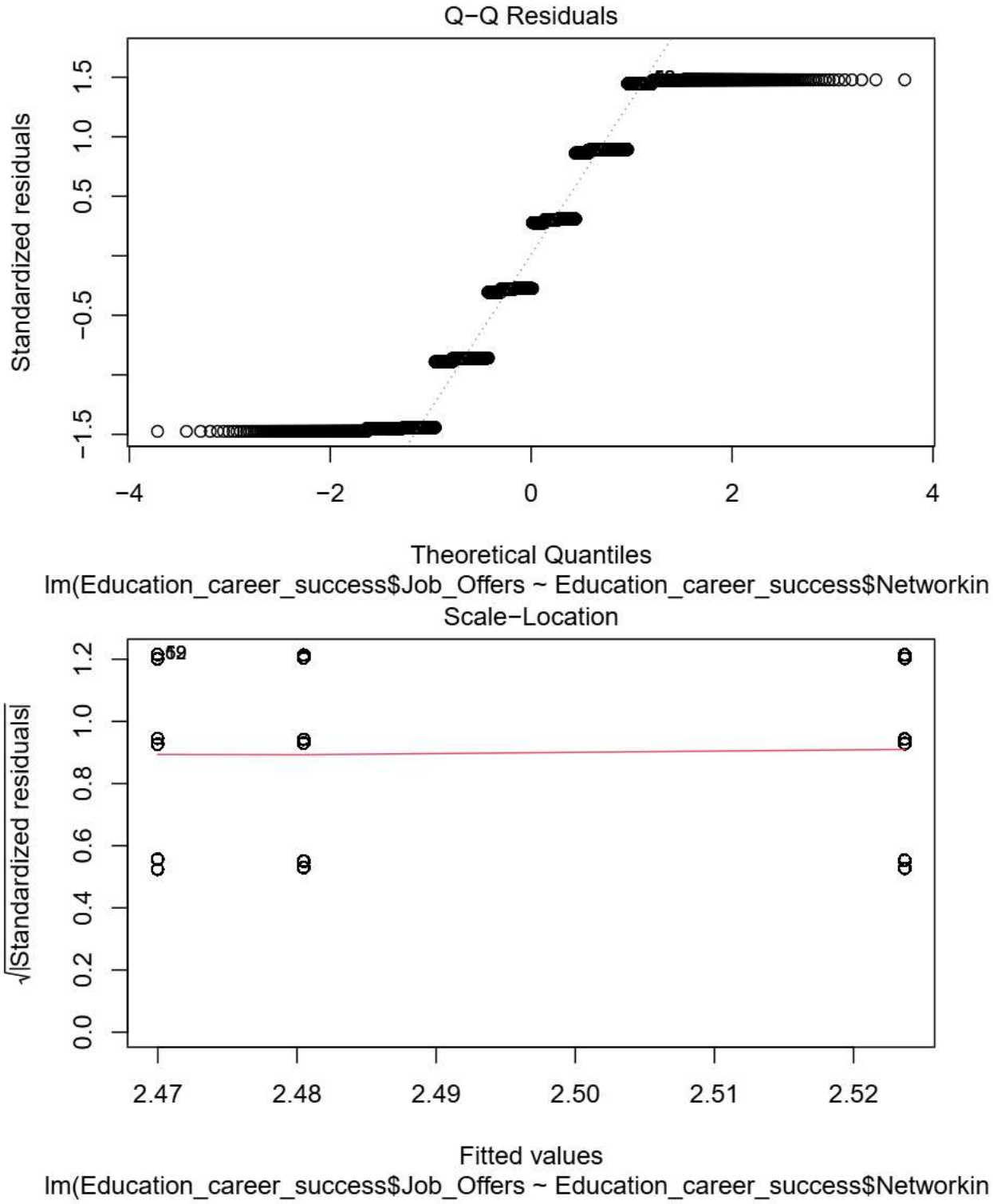
```
## Multiple R-squared:  0.0001773,  Adjusted R-squared:  -0.0002229
## F-statistic: 0.443 on 2 and 4997 DF,  p-value: 0.6421
```

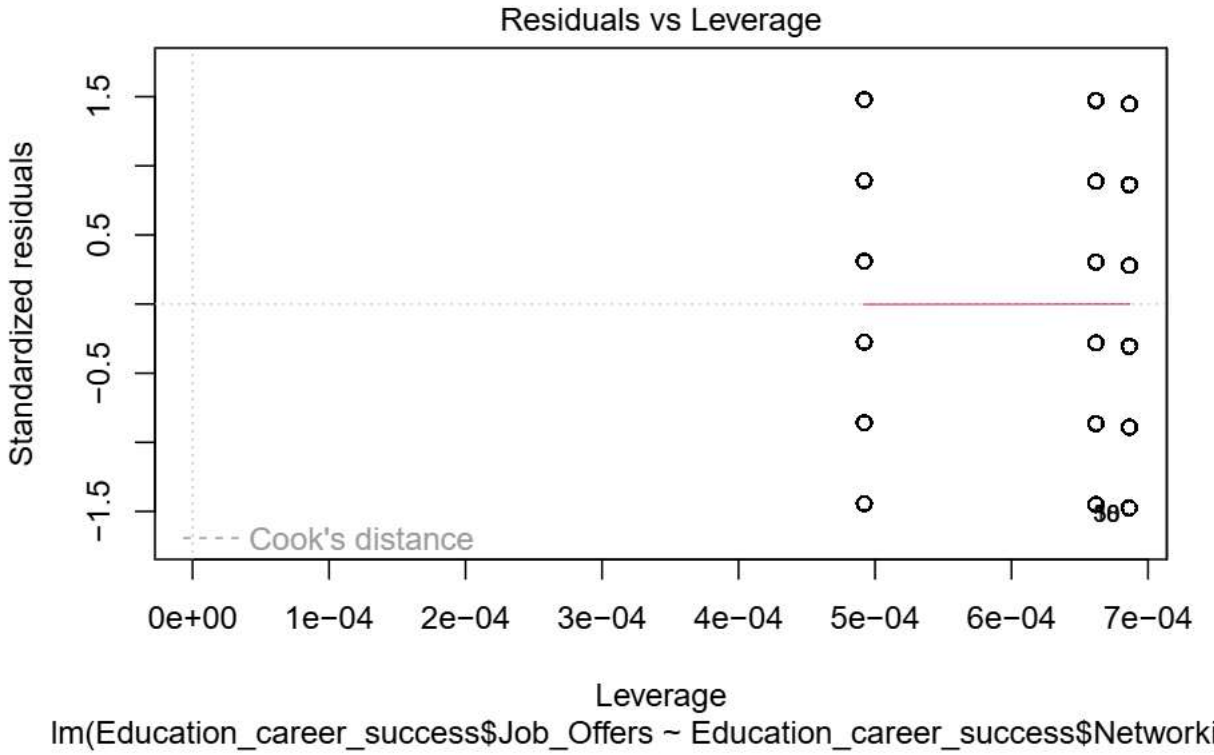
```
par(Edu_money = c(2, 2))
```

```
## Warning in par(Edu_money = c(2, 2)): "Edu_money" is not a graphical parameter
```

```
plot(Edu_money)
```




Q−Q Residuals
Standardized residuals
−1.5
−0.5
0.5
1.0
1.5
−4
−2
0
2
4
Theoretical Quantiles
lm(Education_career_success$Job_Offers ~ Education_career_success$Networkin
Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2.47
2.48
2.49
2.50
2.51
2.52
Fitted values
lm(Education_career_success$Job_Offers ~ Education_career_success$Networkin



QQ residual plot

The plot exhibits an S-shaped curve, meaning the residuals deviate from normality.

The tails (both left and right) deviate significantly from the diagonal line, indicating heavier tails than expected under normality (a common trait with discrete outcome variables).

Your dependent variable is likely discrete and bounded (Job_Offers, ranging from 0–5), which violates the assumptions of linear regression (specifically, normality of errors and homoscedasticity).

Residual Fitted

There's no strong curvature or funnel shape, the residuals cluster tightly and don't vary smoothly with the fitted values. This supports the earlier evidence that the assumptions of homoscedasticity and normality are likely violated.

CONCLUSION

The analysis above indicates that education-related variables, such as University GPA, SAT SCORE, or gender, do not significantly determine the number of job offers a student receives after graduation. This is supported by:

Densely packed and fairly uniform distribution of job offers across the full range (0–5) and the regression line in plot above that is almost horizontal, indicating no meaningful linear relationship between SAT scores, Gpa, gender, age and starting salary.

In contrast, networking-related factors—specifically the Networking Score—show a stronger association with job acquisition success.

The major of study is another variable that slightly positively impact job offers.

The models indicate that students with higher networking scores tend to receive more job offers, even when controlling for other variables.

This aligns with the growing importance of social and professional connections in career success.

REFERENCE Public data source dowloaded from Kaggle

https://www.kaggle.com/datasets/adilshamim8/education-and-career-success?resource=download