DATA 606 Data Project Proposal

2025-04-05

```
getwd()
```

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

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
```

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

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

```r
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
```

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

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

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

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

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

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata
```

```r
library(dplyr)
```

Part 1 - Introduction

We are going to analyze 5000 records of students' educational backgrounds, GPA, SAT scores, and career outcomes.

The relationship between high academic performance and career success will be explored.

We will look at the relationship between job success based on education, identifying key factors influencing salaries, and understanding the role of networking and internships in career growth.

This can be considered an Observational study.

Part 2 - Data

The data set am using in this study can be found on Kaggle.

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

This data contains 5000 observations and 20 variables.

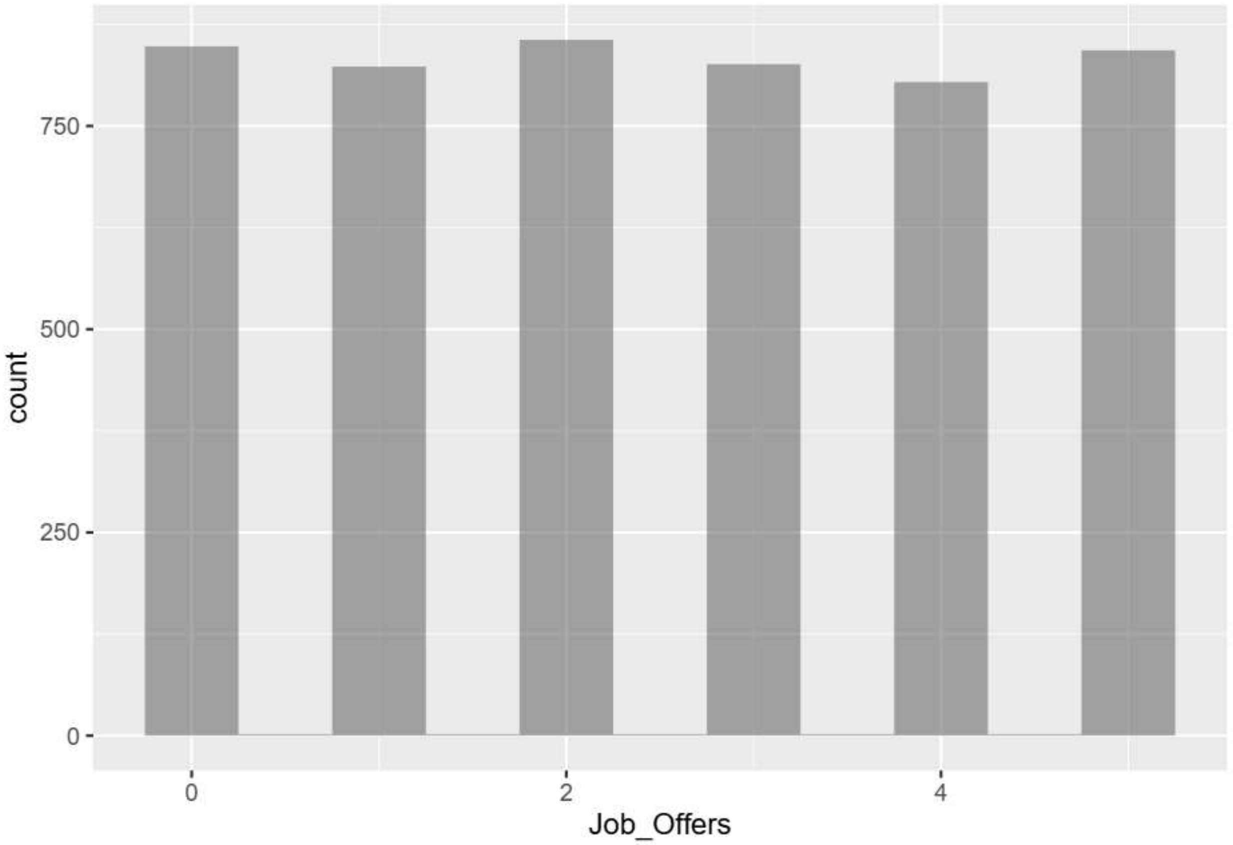
```r
head(Education_career_success)
```

```
##   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
```

Part 3 - Exploratory data analysis

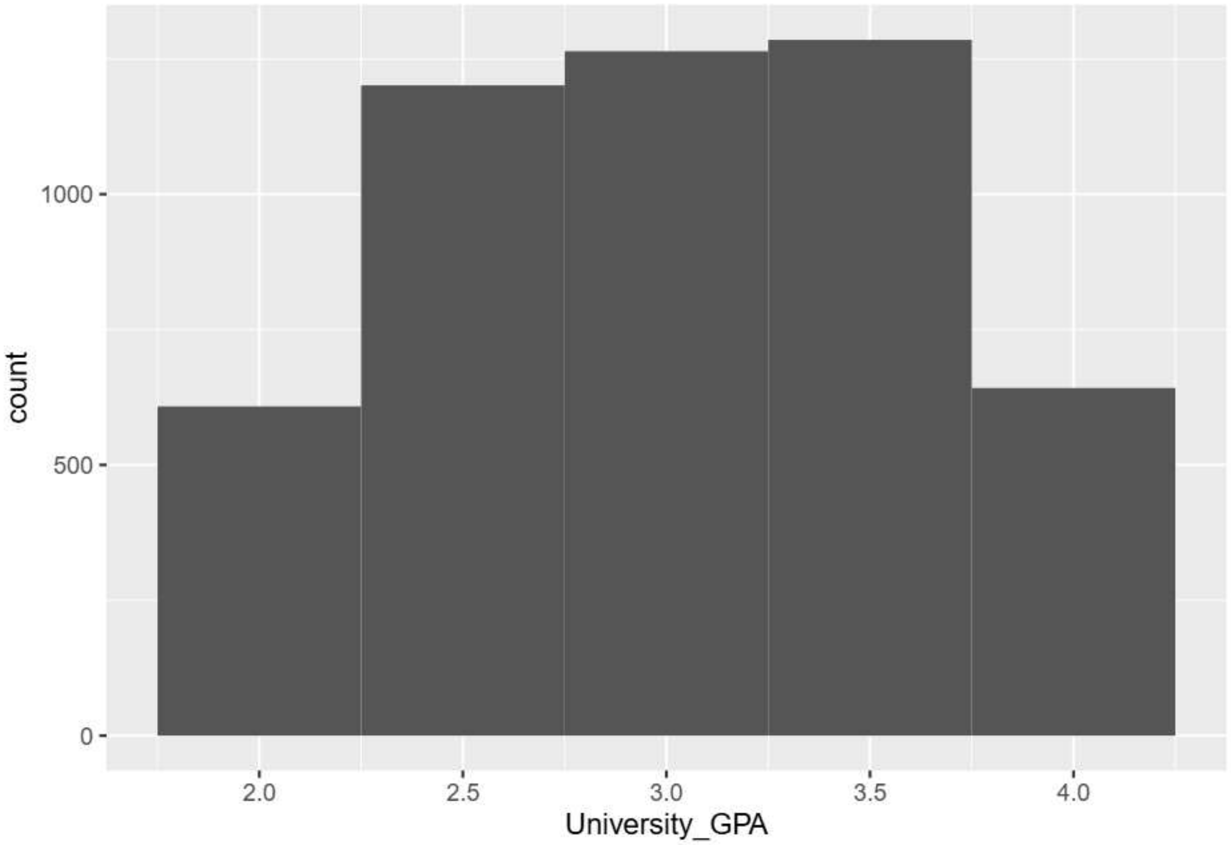
```r
ggplot(Education_career_success, aes(x=Job_Offers, fill=University_GPA)) +
    geom_histogram(binwidth=.5, alpha=.5, position="identity")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



```r
ggplot(Education_career_success, aes(x=University_GPA, fill=Job_Offers)) +
    geom_histogram(binwidth=.5, position="dodge")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

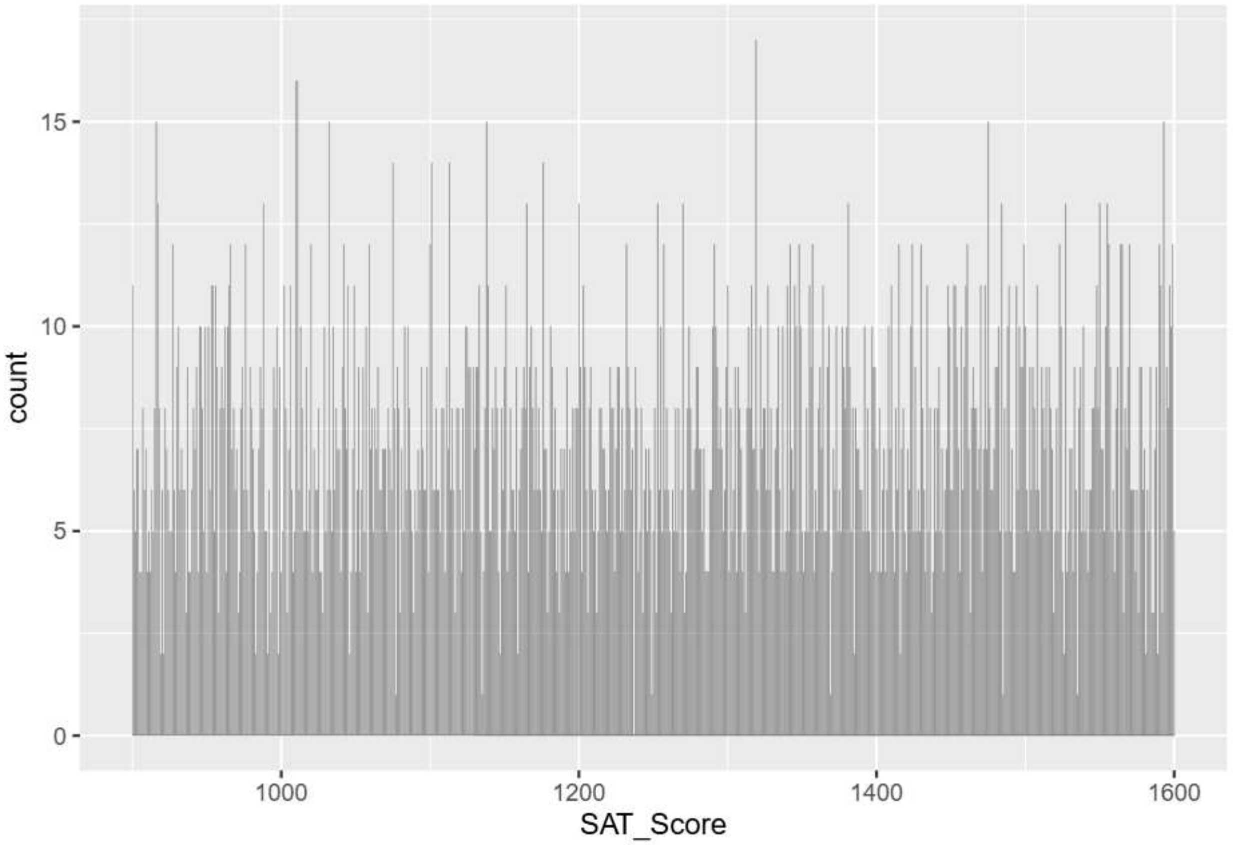


```
  labs(tittle = 'Job offers based on University GPA')
```

```
## $tittle
## [1] "Job offers based on University GPA"
##
## attr(,"class")
## [1] "labels"
```

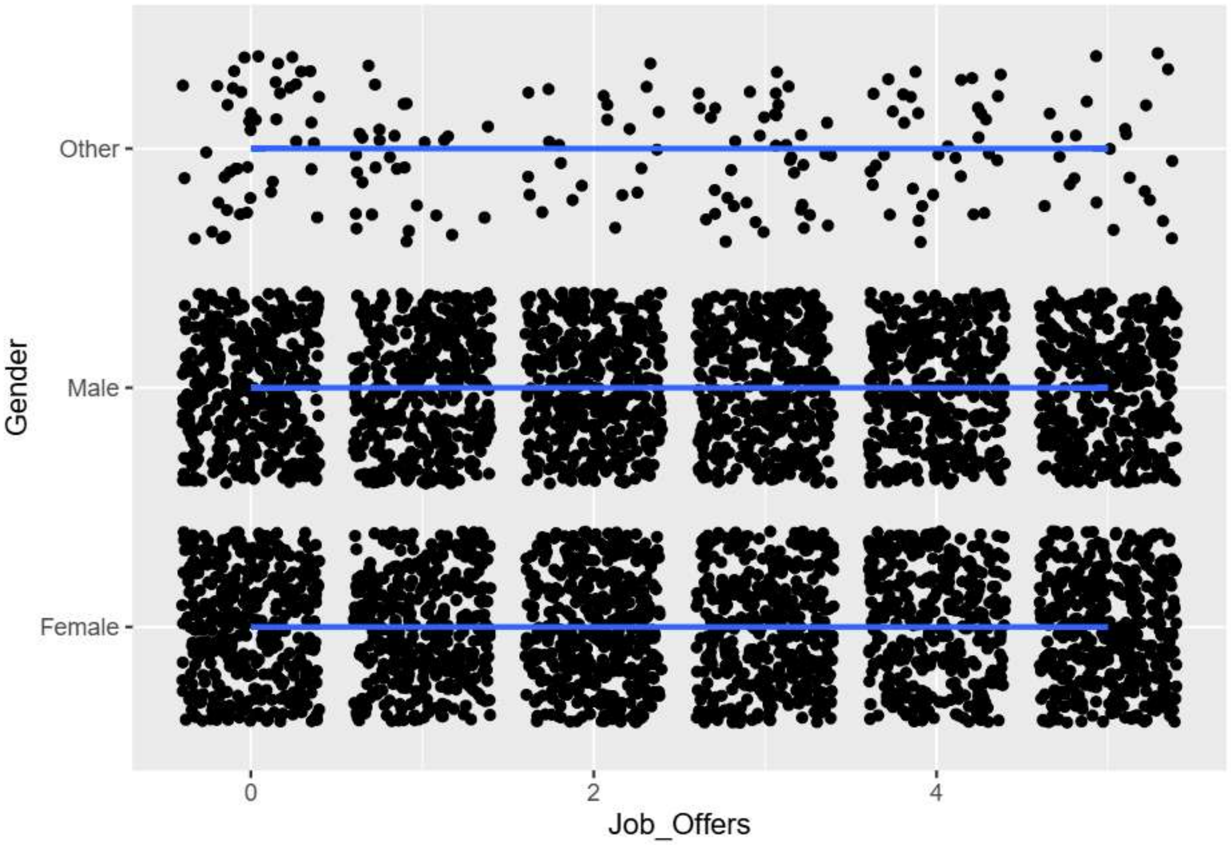
```
ggplot(Education_career_success, aes(x=SAT_Score, fill=Job_Offers)) +  geom_histogram(binwidth=.5, posi
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



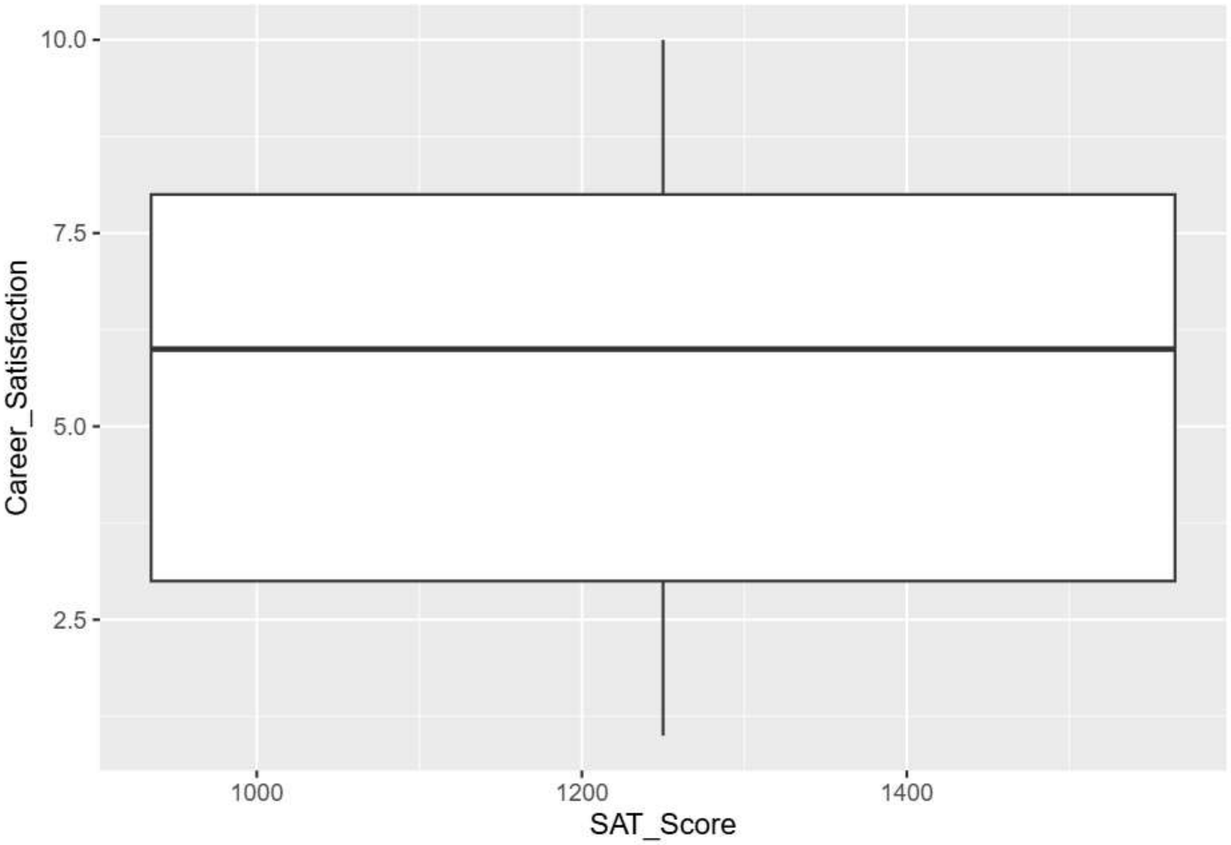
```
ggplot(data = Education_career_success, aes(x = Job_Offers, y = Gender)) +
  geom_jitter() +
  stat_smooth(method = "lm", se = FALSE)
```

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



```
ggplot(Education_career_success, aes(x=SAT_Score, y=Career_Satisfaction)) + geom_boxplot()
```

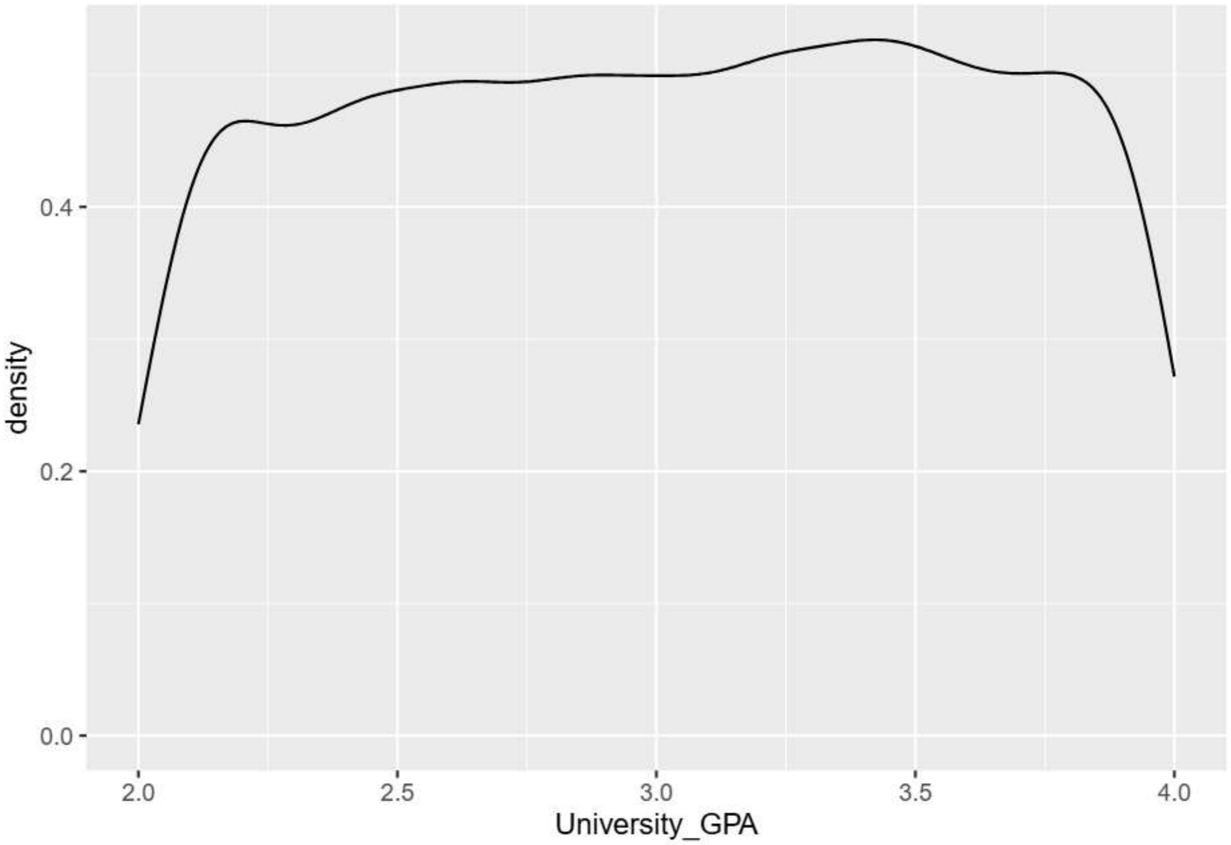
```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```



Part 4 - Inference

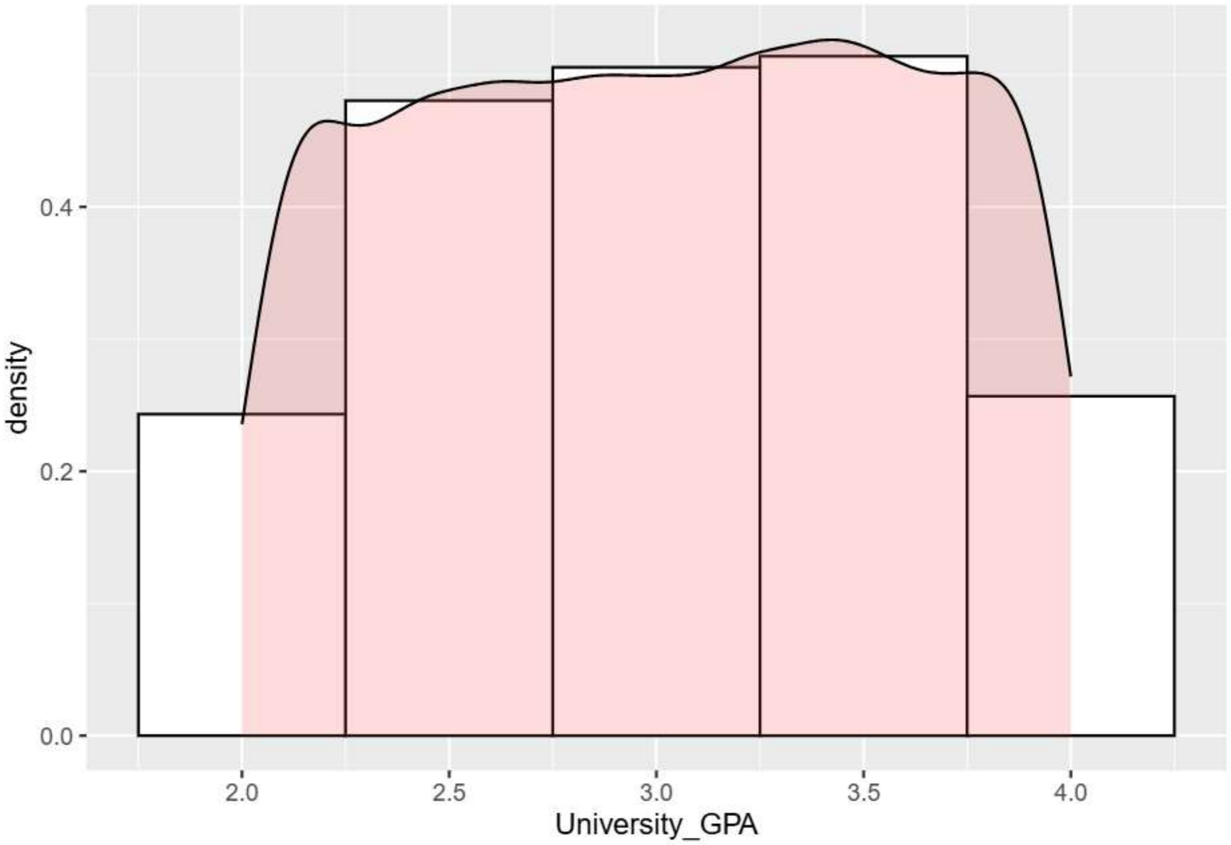
```
ggplot(Education_career_success, aes(x=University_GPA, colour=Job_Offers)) + geom_density()
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



```r
ggplot(Education_career_success, aes(x=University_GPA)) +
    geom_histogram(aes(y=..density..),      # Histogram with density instead of count on y-axis
                   binwidth=.5,
                   colour="black", fill="white") +
    geom_density(alpha=.2, fill="#FF6666")
```

```
## 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.
```



Part 5 - Conclusion

References

Appendix (optional)

Remove this section if you don’t have an appendix