Tuition Paid, Salary Gained: How GPA, Major and Gender Affect Pay

Many young adults have to navigate what's commonly known as the "decade of decisions". These decisions include whether or not to go to college, where to work, who to make life-long relationships with, whether to get married or not, have kids, so on and so forth. Within these decisions, one decision that is often made early for those who go to college is which major to choose. This has a great impact on your future as it will decide who you work with, what work you are doing, what jobs you are able to get in the future, and most importantly, how much money you will make. It can be a daunting decision, but we are here to help you gain some important information on salary earnings, especially as it relates to your GPA, major, and gender.

Looking at the data, there is a clear trend that the better your GPA is, the higher your earnings will be. When running our analysis, we found that for every point increase in GPA, the average increase in salary is $4,800. This can significantly vary, with the increase going as low as $3,300 and as high as $6,300. Regardless, the data clearly shows that better GPAs lead to better earnings.

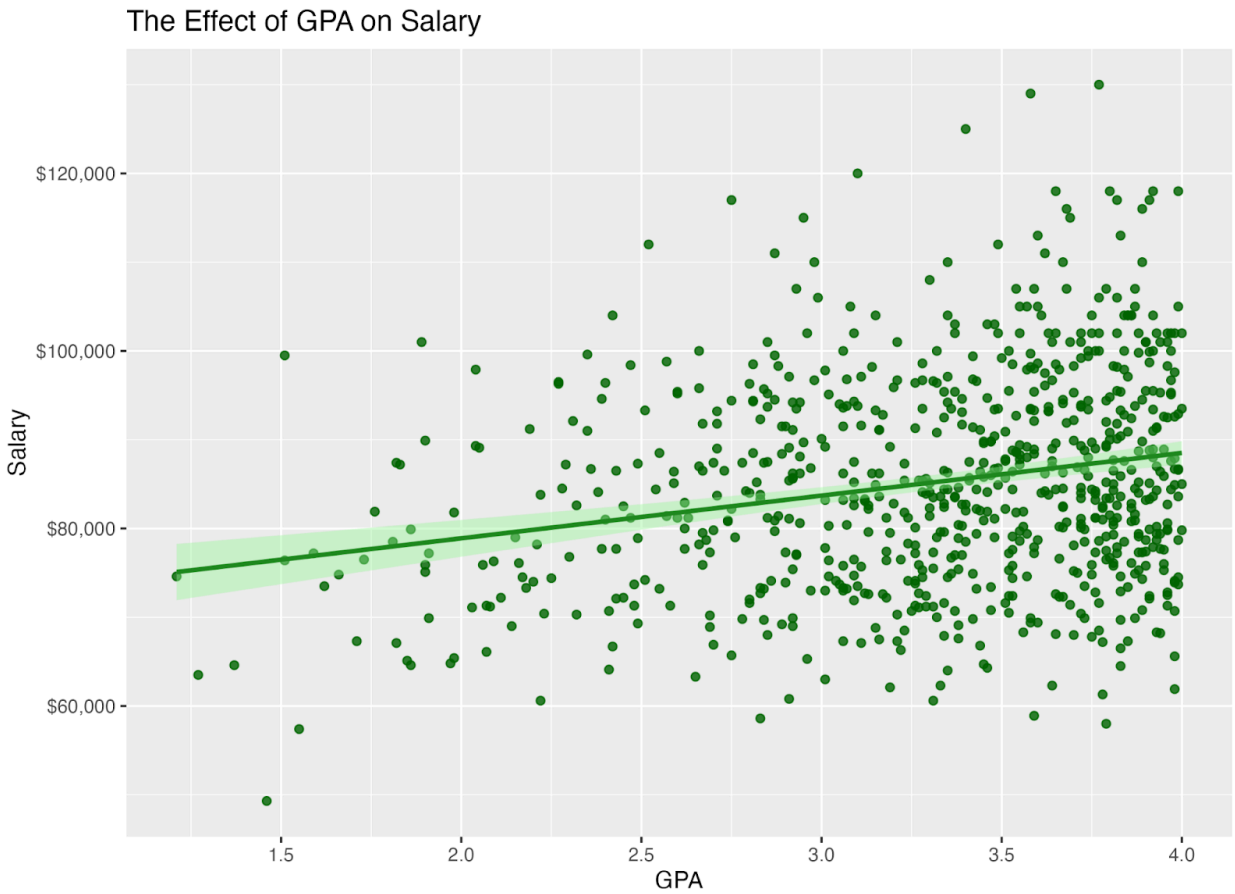


Figure 1: Relationship between GPA and Salary for college students.

Investigating this further, while it makes a lot of sense that a better GPA will net you more money, it is important to consider why this is the case. GPA alone is likely not going to land you a higher paying job, but rather, students who are getting high GPAs are also the ones who are getting internships, admittance to higher education, etc. My advice to students would be that beyond getting good grades, it's important to know where you want your education to get you to. Some internships or schools may require good grades, but it may not be necessary to be a 4.0 student. Getting good grades will lead you to more prestigious opportunities, which will lead you to better paying positions, not necessarily that a straight-A student is guaranteed a high paying job.

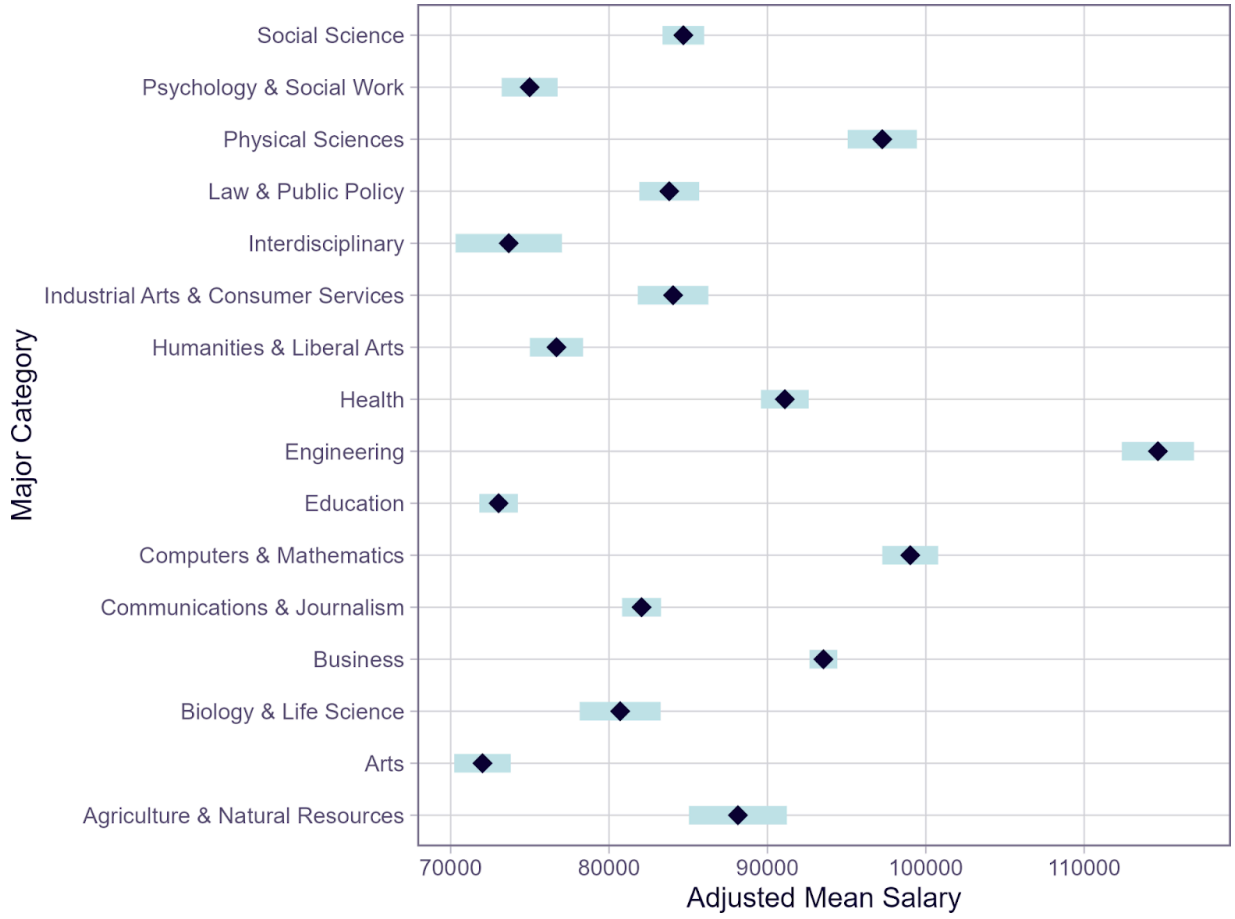


Figure 2: Average Earnings by Major, with Bars Showing the Likely Range of True Earnings

While GPA plays an important role in earnings, what students choose to study also makes a meaningful difference. Even when comparing graduates with similar GPAs, salaries vary noticeably across major categories, with some fields consistently leading to higher earnings five years after graduation than others. Engineering stands out as the highest-paying major category, followed by fields such as Computers and Mathematics, Physical Sciences, Business, and Health-related majors, which tend to

align with industries that place a high value on technical skills, quantitative training, or specialized professional knowledge. In contrast, majors in areas such as Education, Arts, and Psychology and Social Work tend to have lower average salaries, reflecting differences in industry pay structures, funding sources, and career paths rather than a lack of value in these fields. Overall, the data suggests that while GPA matters, a student's choice of major can have a substantial impact on earnings outcomes after college.

Some may wonder while looking at these differences between college majors if there is a difference between the pay of men and women. We investigated this difference and came to find out that more often than not, men earned more money than women do. The difference in pay was largest in interdisciplinary, business, industrial arts, and art majors, while it was the smallest in psychology, agriculture, health, and humanities. These latter four groups were the only ones that may not have a difference in pay at all, while the other 12 groups do have a significant difference in pay.

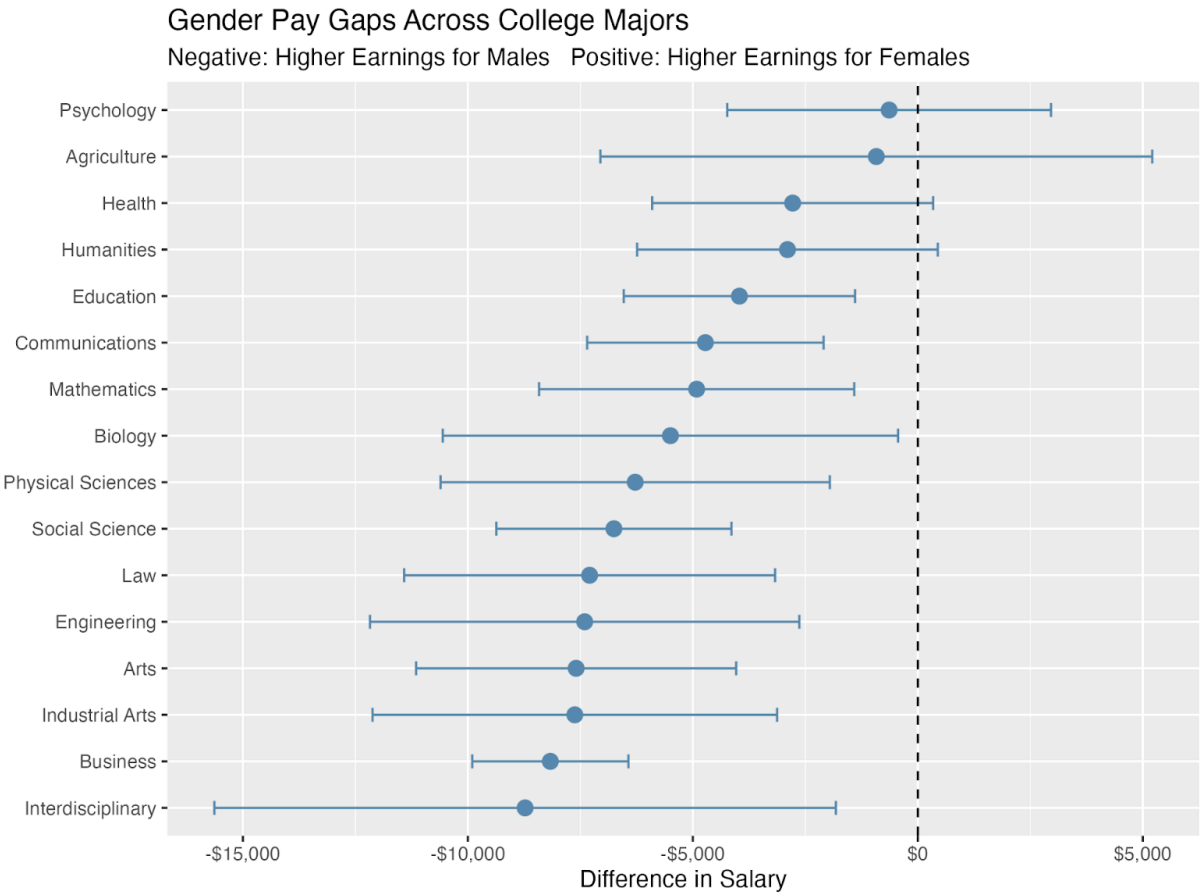


Figure 3: Difference in Earnings between Males and Females among the same Major.

It is hard to identify any meaningful trend in which groups have a difference in pay and which ones don't. Both groups have majors which are technical like health and

business, while also having some more creative majors like arts and humanities. Regardless of finding a trend in which groups pay both genders equally or not, it is undeniable that men get paid more than women. Some factors that may cause this is parental leave having a greater affect on women than men or existing biases which tend to favor men in the workforce and women at home. This should guide you in your decision of which major to pick by helping you get into a field that treat both genders equally or by changing the field you are in to even the playing field.

Overall, the information in this dataset does a reasonably good job of predicting salary, but it is far from perfect. When using GPA, major, and gender to estimate earnings, salary predictions are typically off by about $5,700. To see this in practice, consider a female graduate who majored in Communications and Journalism with a GPA of 2.79. Based on these characteristics alone, the model predicts a salary of about $76,500, while her actual salary was $84,200, a difference of nearly $7,700. This gap highlights how factors not included in the data, such as degree level or job location, can play a major role in determining earnings. While GPA and major clearly matter, this example shows that they are only part of a much larger picture when it comes to real-world salary outcomes.

Choosing a college major and working toward a strong GPA are important decisions that can meaningfully influence earnings after graduation. Our analysis shows that higher GPAs are associated with higher salaries and that certain majors, particularly technical and quantitative fields like Engineering and Computers and Mathematics, tend to lead to higher pay five years after graduation. At the same time, differences in earnings between men and women remain present across many fields, highlighting ongoing disparities in the workforce. However, salary is not determined by GPA or major alone. Even with this information, predictions can still vary by thousands of dollars, reminding us that factors such as location, industry, and degree level all play a role in shaping career outcomes. As students navigate their own “decade of decisions,” the goal should not simply be to chase the highest-paying major or achieve a perfect GPA, but to make informed choices while recognizing that long-term success depends on preparation, opportunity, and personal direction.

Documentation

In this section, we will cover the methods used in our analysis for other data analysts to inspect and reuse. The mathematical model we used was a multivariate normal model in the following form:

$$y \sim N(X\beta,\ \sigma^2 I)$$

Where y is the salary in dollars and X is a matrix which included the intercept, GPA, Major, and Gender, the latter two which were encoded using dummy variables. All of our models are the same general linear models, with different choices of X. In section 2, we did regression with GPA alone and we used GPA, major, and the interaction between them. In section 3, we used GPA, major, and gender. In section 4, we used GPA, major, gender, and the interaction between major and sex. Each of these models best answered each question we were trying to understand and were appropriate for each context.

Going through the validation of our model, we made an added variable plot looking at GPA, and we can confidently say that there is no problem in linearity between GPA and salary. Given what we know about our dataset, we have no reason to believe that the salary of one person will affect the salary of another person, so we will assume independence in our observations. We looked at the histogram of the residuals and ran a KS Test, which both gave us no reason to doubt that the residuals are normally distributed. We also looked at a fitted vs residuals plot and a BP test which gave us the assurance that equal variance is true.

In addition to validating model assumptions, we evaluated how well the model performs when predicting new observations. In-sample model fit can sometimes give overly optimistic results, so we used 20-fold cross-validation to assess predictive performance. In each fold, the model was trained on 19 subsets of the data and tested on the remaining subset. We calculated the root mean squared error (RMSE) for each fold and averaged them to obtain an overall prediction error of 5,717.57. This value represents the typical magnitude of prediction error when applying the model to new data. The consistency of RMSE across folds indicates that the model's predictive performance is stable, but also highlights that a meaningful amount of salary variation remains unexplained.

In Section 2, we looked at two different models to answer questions about the effect of GPA on salary and how that effect differs between majors. Our first model looked at GPA alone and found the linear relationship between GPA and salary. We also found a 95% confidence interval of GPA which explains the expected gain in salary for every point increase in GPA. We also added a section at the end which looks at the sample size for each group to best understand which estimates are certain and which estimates are less certain.

For Section 3, we ran a linear model with GPA, major, and gender to determine whether salary differs across majors after adjusting for GPA and sex. Major and gender

were treated as categorical variables using dummy coding. We then conducted an F-test to evaluate whether the major category as a whole contributes significantly to predicting salary. The test produced an extremely small p-value, which gives strong evidence that salary differs across majors, even after accounting for GPA and gender.

To better understand which majors have higher or lower salaries, we calculated the adjusted (marginal) mean salaries for each major category. These adjusted means represent the expected salary for each major while holding GPA and gender constant, allowing for fair comparisons across fields. From these results, we found that Engineering has the highest adjusted mean salary. Other majors with relatively high salaries include Computers & Mathematics, Physical Sciences, Business, and Health. Majors such as Education, Arts, and Psychology & Social Work have lower adjusted mean salaries. We plotted these adjusted means along with 95% confidence intervals to visually display the differences and associated uncertainty.

For Section 4, we ran a linear model with GPA, major, gender, and the interaction between major and gender. We then looked at the marginal means for gender in order to understand the difference in pay between men and women in each major category. We got an estimate for this, along with a 95% confidence interval of the difference which we plotted to show where the difference could be.

For Section 5, we evaluated how well our model predicts salary using cross-validation. We implemented 20-fold cross-validation by randomly splitting the dataset into 20 groups, repeatedly fitting the model on 19 groups and testing it on the remaining group. For each fold, we computed the root mean squared error (RMSE), which measures the typical size of prediction error in dollars. The average RMSE across all folds was 5,717.57, meaning that when predicting salary for new observations, the model is typically off by about $5,700.

To further illustrate prediction performance, we examined an individual out-of-sample prediction from one of the folds. In this case, we considered a female graduate who majored in Communications and Journalism with a GPA of 2.79 and an observed salary of $84,200. Using a model trained without this observation, the predicted salary was $76,512.96, resulting in a prediction error of $7,687.03. This difference is consistent with the magnitude of error reflected in the cross-validated RMSE and demonstrates that while GPA, major, and gender explain some variation in salary, substantial unexplained variability remains.

These results suggest that additional variables not included in the dataset, such as degree level or job location, likely contribute meaningfully to salary outcomes. Including such variables in future models would likely reduce prediction error and improve overall predictive performance.

College Education

Cooper Riggs & Stockton Nelson

Setup

```
#Downloading Libraries
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   1.6.0
v ggplot2   4.0.1     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 becor
```

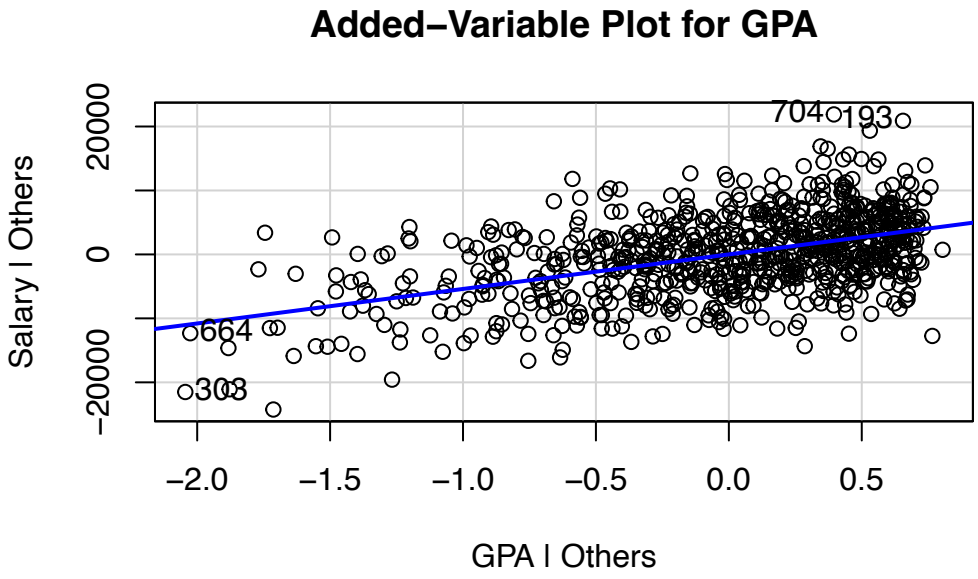
```
#Bringing in Data
salary <- vroom::vroom("Salary.csv") |>
  janitor::clean_names() |>
  rename(major = major_category)
```

```
Rows: 773 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): MajorCategory, Sex
dbl (2): Salary, GPA

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Validation

Linearity

```
#Making General Linear Model
glm <- lm(salary ~ ., data = salary)

#Checking Linearity of GPA
car::avPlot(glm, "gpa",
            xlab = "GPA | Others",
            ylab = "Salary | Others",
            main = "Added-Variable Plot for GPA")
```



Answer: It looks like GPA has a nice, linear trend. There are no concerns here.

Independence

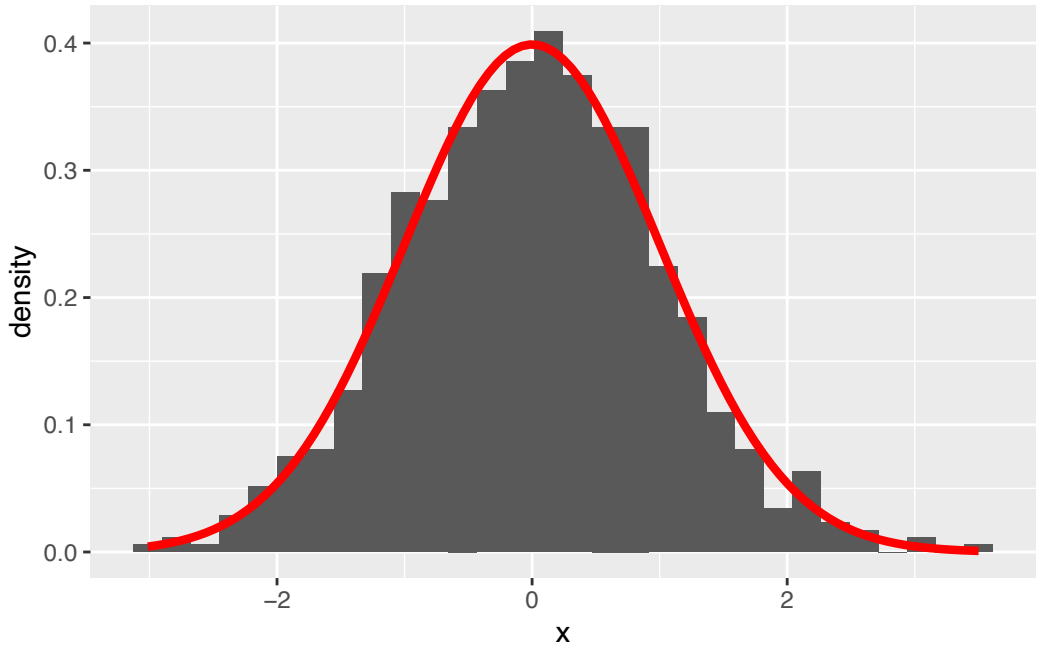
Answer: We have no reason to believe that the salary of one person will affect the salary of another person.

Normality

```
stresids <- MASS::stdres(glm)
ggplot(data = data.frame(x = stresids), mapping = aes(x = x)) +
  geom_histogram(aes(y = after_stat(density))) +
  stat_function(fun = dnorm, color = "red", size = 1.5)
```

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

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




```
#Running a KS-Test
ks.test(x = stresids, "pnorm", mean = 0, sd = 1)
```

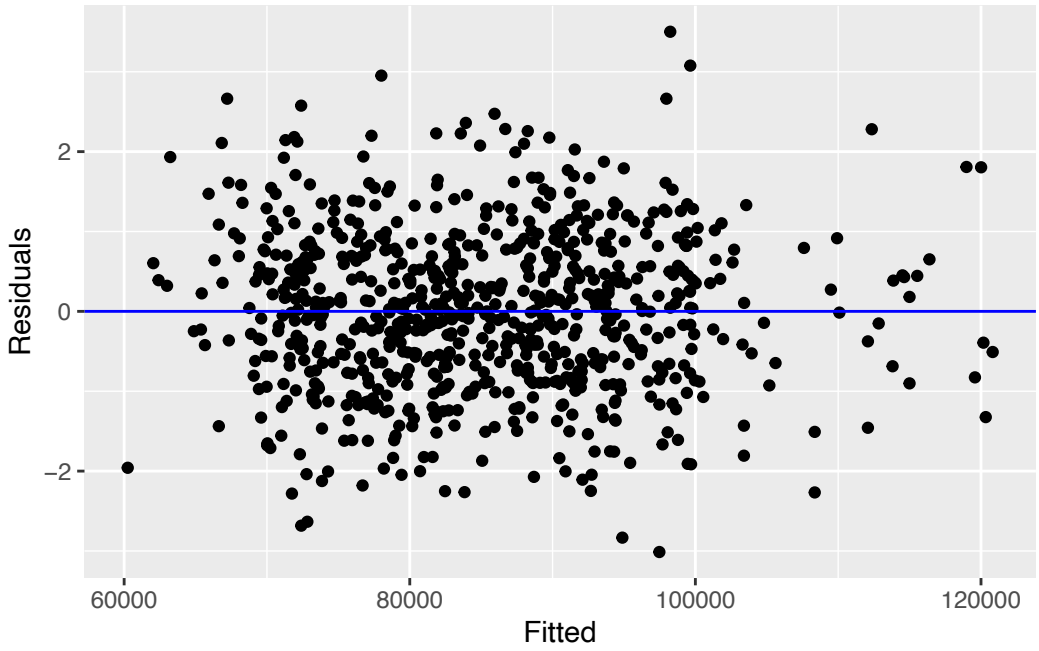
```
Warning in ks.test.default(x = stresids, "pnorm", mean = 0, sd = 1): ties
should not be present for the one-sample Kolmogorov-Smirnov test
```

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  stresids
D = 0.016463, p-value = 0.9848
alternative hypothesis: two-sided
```

Answer: There is no concern whatsoever with normality when looking at the histogram. We also are not concerned when running the KS Test.

Equal Variance

```
#Checking Equal Variance Assumption
ggplot(data = data.frame(fit = fitted(glm), res = stresids),
       aes(x = fit, y = res)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "blue") +
  labs(x = "Fitted",
       y = "Residuals")
```



```
#Running a BP-Test
lmtest::bptest(glm)
```

```
    studentized Breusch-Pagan test

data:  glm
BP = 15.154, df = 17, p-value = 0.5844
```

Answer: Looking at the Fitted Vs Residuals Plot, we aren't concerned about equal variance being an issue. Looking at a BP test as well, we feel good about assuming equal variance.

Problem 2

Impact of GPA on Salary

```
#Running Linear Model with GPA
lmgpa <- lm(salary ~ gpa, data = salary)

#Running Confidence Interval on GPA
confint(lmgpa, "gpa", level = 0.95)
```

```
       2.5 %   97.5 %
gpa 3355.021 6265.605
```

```
#Making Scatterplot of GPA and Salary
ggplot(data = salary, mapping = aes(x = gpa, y = salary)) +
  geom_point(color = "darkgreen", alpha = 0.8) +
  geom_smooth(method = lm, se = T, color = "forestgreen", fill = "palegreen") +
  scale_x_continuous(breaks = c(1.5, 2.0, 2.5, 3.0, 3.5, 4.0),
                     labels = function(x) sprintf("%.1f", x)) +
  scale_y_continuous(labels = scales::dollar_format()) +
  labs(x = "GPA",
       y = "Salary",
       title = "The Effect of GPA on Salary",
       caption = "Figure 1: Relationship between GPA and Salary for college students.") +
  theme(plot.caption = element_text(hjust = 0))
```

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

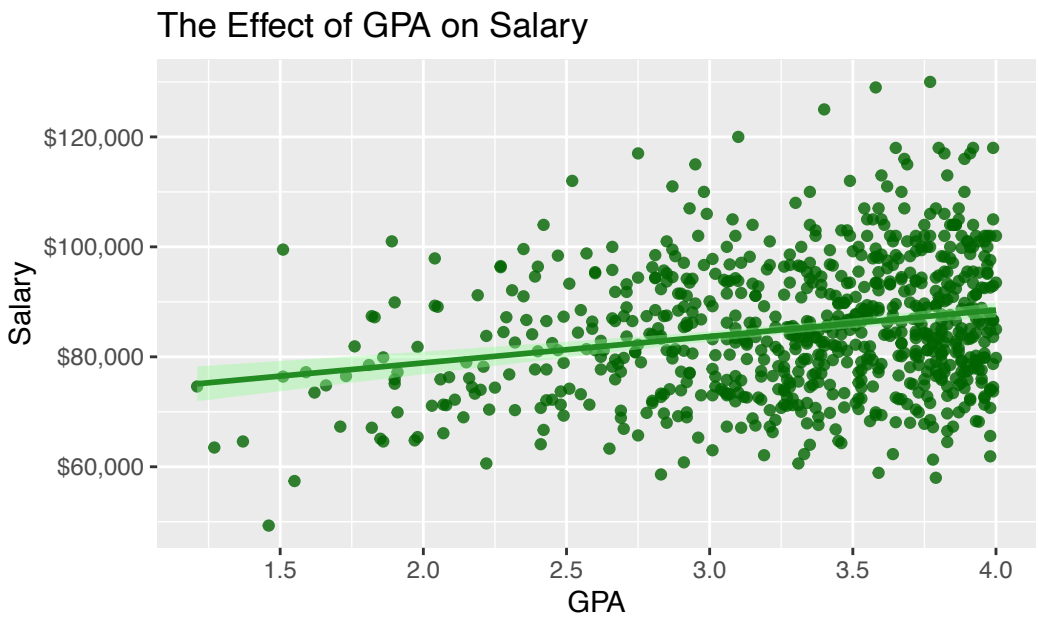


Figure 1: Relationship between GPA and Salary for college students.

```
ggsave("GPA_salary_plot.png", width = 8, height = 6, dpi = 300)
```

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

Answer: A greater GPA has a positive impact on your salary. For every one point your GPA goes up, we expect your salary to go up somewhere between $3,300 and $6,300 dollars. While there is a positive impact on your salary, I do believe that GPA does positively increase your salary, I believe that this can mainly be explained by students with higher GPAs have a greater opportunity in what internships and graduate schools they are able to attend.

Impact of GPA across Majors

```
#Running Linear Model with GPA and Major Interaction
lmgm <- lm(salary ~ gpa * major, data = salary)

#Running Confidence Interval on GPA
fit <- lm(salary ~ gpa * major, data = salary)
confint(fit, level = 0.95)
```

```
                                                  2.5 %     97.5 %
(Intercept)                                    53592.963 119714.192
gpa                                            -8548.648  10002.478
majorArts                                     -63625.150   7686.972
majorBiology & Life Science                   -54643.213  19942.552
majorBusiness                                 -48276.175  18726.795
majorCommunications & Journalism              -58597.829   9235.248
majorComputers & Mathematics                  -27100.954  42270.565
majorEducation                                -60697.141   7027.321
majorEngineering                              -36785.955  39137.938
majorHealth                                   -47479.507  21548.139
majorHumanities & Liberal Arts                -50883.889  17992.484
majorIndustrial Arts & Consumer Services      -69175.059   2555.377
majorInterdisciplinary                        -48345.724  39204.386
majorLaw & Public Policy                      -59972.518  11441.352
majorPhysical Sciences                        -58188.766  19912.670
majorPsychology & Social Work                 -65161.549   6405.796
majorSocial Science                           -60298.203   7672.858
gpa:majorArts                                  -6830.097  13297.525
gpa:majorBiology & Life Science                -8006.184  13405.381
gpa:majorBusiness                              -3592.935  15241.400
gpa:majorCommunications & Journalism           -4536.557  14541.579
gpa:majorComputers & Mathematics               -9283.012  10343.390
gpa:majorEducation                             -6553.671  12496.701
gpa:majorEngineering                           -3711.366  17792.162
gpa:majorHealth                                -5487.789  14029.624
gpa:majorHumanities & Liberal Arts             -8589.110  10841.269
gpa:majorIndustrial Arts & Consumer Services   -1382.056  18988.775
gpa:majorInterdisciplinary                    -15803.462   9212.260
gpa:majorLaw & Public Policy                   -4704.117  15448.755
gpa:majorPhysical Sciences                     -2909.962  19370.749
gpa:majorPsychology & Social Work              -5703.912  14552.784
gpa:majorSocial Science                        -2992.709  16162.242
```

```
salary |>
  group_by(major) |>
  summarize(n = n())
```

```
# A tibble: 16 x 2
   major                                n
   <chr>                            <int>
 1 Agriculture & Natural Resources     13
```

```
 2 Arts                                    39
 3 Biology & Life Science                  19
 4 Business                               161
 5 Communications & Journalism             84
 6 Computers & Mathematics                 40
 7 Education                               85
 8 Engineering                             24
 9 Health                                  55
10 Humanities & Liberal Arts               44
11 Industrial Arts & Consumer Services     25
12 Interdisciplinary                       11
13 Law & Public Policy                     35
14 Physical Sciences                       26
15 Psychology & Social Work                40
16 Social Science                          72
```

Answer: Looking at the 95% confidence interval for each of these interactions, it's safe to conclude that the effect for GPA is the same, regardless of your major. This could be due to the small sample sizes for each major, but with what data we got, all of these confidence intervals return intervals that contain 0, thus showing that there isn't an interaction.

Problem 3

Differences in Salary Across Major

```r
# Fit linear model to evaluate effect of major on salary
# Model includes GPA and Sex as controls
sal_lm <- lm(salary ~ gpa + sex + major, data = salary)

# Perform overall F-test to determine whether Major contributes significantly to predicting S
anova(sal_lm)
```

```
Analysis of Variance Table

Response: salary
           Df     Sum Sq    Mean Sq F value    Pr(>F)
gpa         1 5.8504e+09 5850439627  182.18 < 2.2e-16 ***
sex         1 8.2766e+09 8276551378  257.73 < 2.2e-16 ***
major      15 7.4614e+10 4974287278  154.90 < 2.2e-16 ***
Residuals 755 2.4245e+10   32112868
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

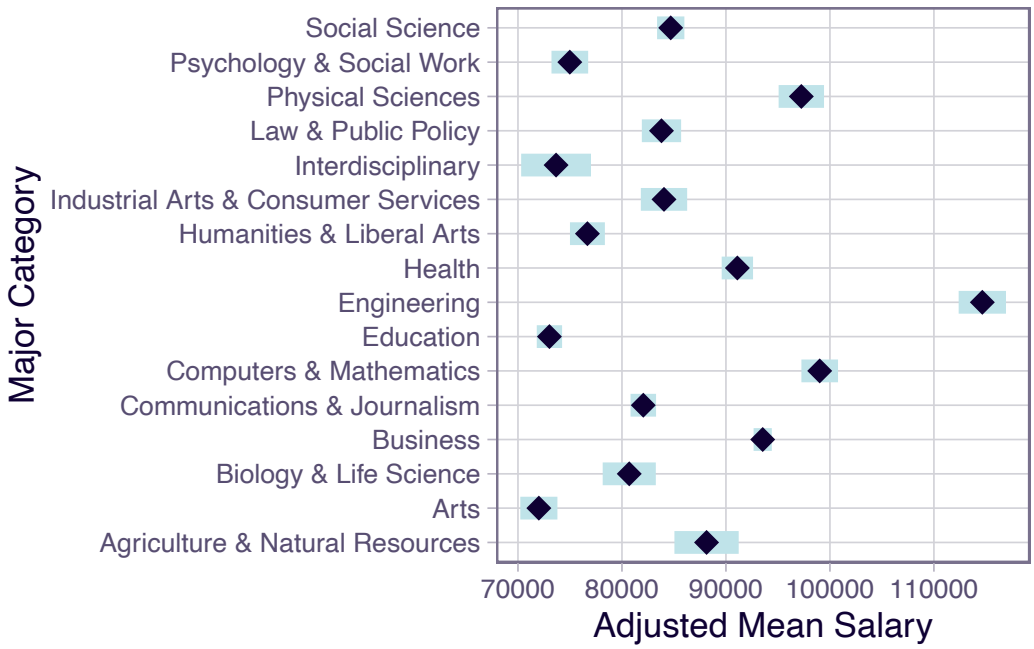
```r
# Load emmeans package to compute adjusted (marginal) means
library(emmeans)
```

```
Welcome to emmeans.
Caution: You lose important information if you filter this package's results.
See '? untidy'
```

```r
# Compute adjusted mean salary for each major category
# GPA and Sex are held constant for fair comparison
major_means <- emmeans(sal_lm, ~ major)

# Obtain 95% confidence intervals for adjusted means
confint(major_means)
```

```
 major                              emmean   SE  df lower.CL upper.CL
 Agriculture & Natural Resources     88135 1570 755    85045    91225
 Arts                                72009  908 755    70226    73792
 Biology & Life Science              80700 1300 755    78147    83254
 Business                            93534  447 755    92657    94411
 Communications & Journalism         82051  624 755    80826    83277
 Computers & Mathematics             99016  897 755    97256   100777
 Education                           73024  619 755    71808    74240
 Engineering                        114654 1160 755   112378   116930
 Health                              91098  768 755    89591    92604
 Humanities & Liberal Arts           76678  855 755    75001    78356
 Industrial Arts & Consumer Services 84042 1130 755    81816    86269
 Interdisciplinary                   73665 1710 755    70308    77021
 Law & Public Policy                 83800  962 755    81910    85689
 Physical Sciences                   97249 1110 755    95067    99431
 Psychology & Social Work            74986  898 755    73224    76748
 Social Science                      84689  669 755    83377    86002

Results are averaged over the levels of: sex
Confidence level used: 0.95
```

```r
plot(major_means,
     ylab = "Major Category",
     xlab = "Adjusted Mean Salary")
```




```
ggsave("major_salary_plot.png", width = 8, height = 6, dpi = 300)
```

Answers:

Through running an F-test on the the linear model (Salary ~ GPA + Sex + Major) we get an extremely low p-value for major, suggesting that there is sufficient evidence to reject the null hypothesis. This means that there is a difference in salaries across majors. Through further inspection by taking the adjusted means of salaries for each major category we find that Engineering majors have a significantly higher salary. Other majors with higher salaries include Computers & Mathematics, Physical Sciences, Business, and Health.

Problem 4

Differences in Salary Across Sex

```
#Running Linear Model with Major and Sex Interaction
lmms <- lm(salary ~ gpa + major * sex, data = salary)

#Finding Marginal Means
emm <- emmeans::emmeans(lmms, ~ sex | major)
```

```r
#Comparing Marginal Means Between Sex
confint(pairs(emm), level = 0.95)
```

```
major = Agriculture & Natural Resources:
 contrast estimate   SE  df lower.CL upper.CL
 F - M        -922 3120 740    -7054     5210

major = Arts:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -7594 1810 740   -11151    -4037

major = Biology & Life Science:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -5498 2580 740   -10558     -438

major = Business:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -8167  884 740    -9903    -6431

major = Communications & Journalism:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -4721 1340 740    -7349    -2093

major = Computers & Mathematics:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -4915 1780 740    -8417    -1413

major = Education:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -3965 1310 740    -6535    -1394

major = Engineering:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -7404 2430 740   -12175    -2633

major = Health:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -2781 1590 740    -5904      341

major = Humanities & Liberal Arts:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -2897 1700 740    -6238      445
```

```
major = Industrial Arts & Consumer Services:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -7623 2290 740   -12118    -3127

major = Interdisciplinary:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -8728 3520 740   -15634    -1821

major = Law & Public Policy:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -7294 2100 740   -11416    -3172

major = Physical Sciences:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -6281 2200 740   -10606    -1955

major = Psychology & Social Work:
 contrast estimate   SE  df lower.CL upper.CL
 F - M        -637 1830 740    -4235     2961

major = Social Science:
 contrast estimate   SE  df lower.CL upper.CL
 F - M       -6754 1330 740    -9367    -4141

Confidence level used: 0.95
```

```
#Making a Graph of the Confidence Interval of the Differences
diffs <- pairs(emm) |> confint() |> as.data.frame() |>
  rename(
    Major = major,
    Difference = estimate,
    Lower = lower.CL,
    Upper = upper.CL
  ) |>
  mutate(Major = recode(Major,
                        "Psychology & Social Work" = "Psychology",
                        "Agriculture & Natural Resources" = "Agriculture",
                        "Humanities & Liberal Arts" = "Humanities",
                        "Communications & Journalism" = "Communications",
                        "Computers & Mathematics" = "Mathematics",
                        "Biology & Life Science" = "Biology",
```

```
                          "Law & Public Policy" = "Law",
                          "Industrial Arts & Consumer Services" = "Industrial Arts"))
ggplot(diffs, aes(x = Difference, y = reorder(Major, Difference))) +
  geom_point(color = "#598BAF", size = 3) +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.3, color = "#598BAF") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "black") +
  scale_x_continuous(labels = scales::dollar_format()) +
  labs(
    x = "Difference in Salary",
    y = " ",
    title = "Gender Pay Gaps Across College Majors",
    subtitle = "Negative: Higher Earnings for Males   Positive: Higher Earnings for Females"
    caption = "Figure 3: Difference in Earnings between Males and Females among the same Majo
  theme(plot.caption = element_text(hjust = 0))
```

```
Warning: `geom_errorbarh()` was deprecated in ggplot2 4.0.0.
i Please use the `orientation` argument of `geom_errorbar()` instead.
```

```
`height` was translated to `width`.
```

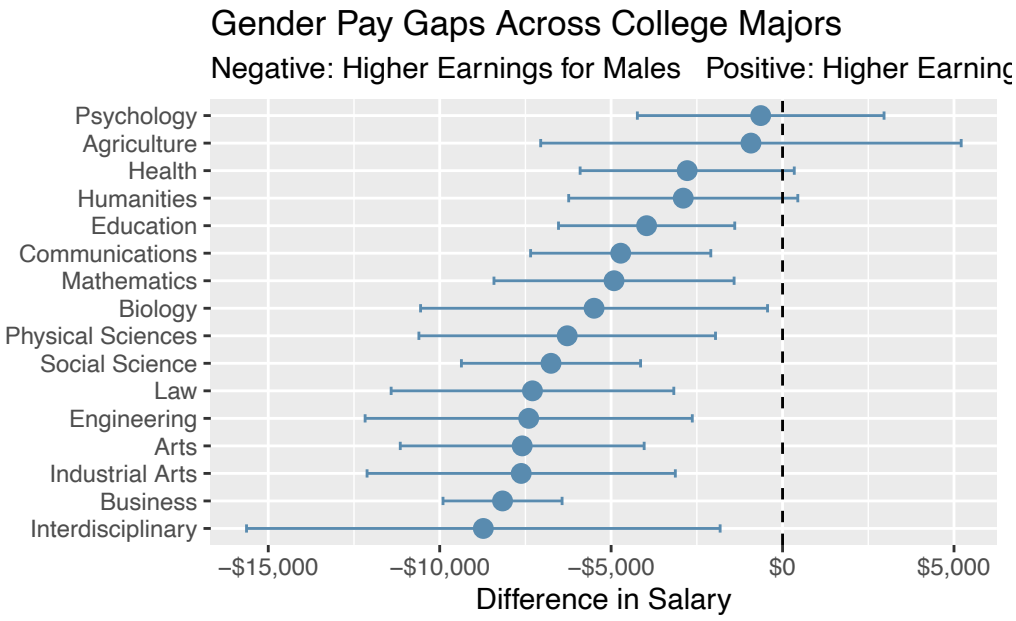


Figure 3: Difference in Earnings between Males and Females among the sa

```r
#Saving the Plot
ggsave("Gender_Pay_Gap.png", width = 8, height = 6, dpi = 300)
```

```
`height` was translated to `width`.
```

Answers: There is a difference between the earnings for males and females between college major. There are some majors where there is no difference, but there are also some majors where there's a stark difference and males earn most of the income.

Problem 5

How Well do the Variables Explain Salary

```r
# Set seed for reproducibility so fold assignment is consistent
set.seed(123)

# Function to compute RMSE for a given fold
cross_validate <- function(fold_num) {

  # Split data into training and testing sets
  train_data <- salary[folds != fold_num, ]
  test_data  <- salary[folds == fold_num, ]

  # Fit linear model on training data only
  fit <- lm(
    salary ~ gpa + sex + major,
    data = train_data
  )
  # Generate predictions for test data
  preds <- predict(fit, newdata = test_data)

  # Compute RMSE for this fold
  rmse <- sqrt(mean((test_data$salary - preds)^2))

  return(rmse)
}

# Define number of folds for cross-validation
K <- 20
```

```r
# Randomly assign each observation to one of K folds
folds <- rep(1:K, length = nrow(salary)) %>%
  sample()

# Apply cross-validation across all folds
rmse_results <- lapply(1:K, FUN = cross_validate) %>%
  unlist()
# Compute average RMSE across folds
mean(rmse_results)
```

```
[1] 5717.572
```

```r
# Select one fold to demonstrate an out-of-sample prediction
fold_num <- 3

# Recreate training and testing split for that fold
train_data <- salary[folds != fold_num, ]
test_data  <- salary[folds == fold_num, ]

# Fit model using training data only
fit <- lm(salary ~ gpa + sex + major, data = train_data)

# Select one observation from the test set
example <- test_data[1, ]

# Generate prediction for this observation
predicted_salary <- predict(fit, newdata = example)

# Extract actual salary and compute prediction error
actual_salary <- example$salary
error <- actual_salary - predicted_salary

# Display actual value, predicted value, and error
print(c(actual_salary, predicted_salary, error))
```

```
                1         1
84200.000 76512.957  7687.043
```

Answers:

The variables in the data set do a pretty good job at predicting salary. After RMSE cross-validation, the mean RMSE across the folds is 5717.57. This means that on average, the model can predict a given person's salary with an error of about $5,717.57. That's okay for predicting, but definitely not great. A couple of factors that are missing that we believe would also contribute to salary are degree level and location of the job. Adding these factors would likely lower the average RMSE. A specific example that shows how this model performs is if you had a female that majored in Communication and Journalism with a 2.79 GPA and a salary of $84,200. The model would predict someone with these characteristics to have a salary of $76,512.96, this is off by $7,687.03. We would say that the difference between $84,200 and $76,512 in salary is fairly significant.