Regression Project

AUTHOR

Group 8

Introduction

```r
library(causaldata)
```

```
Warning: package 'causaldata' was built under R version 4.3.2
```

```r
library(datasets)
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(ggplot2)
library(emmeans)
head(scorecard)
```

```
  unitid                          inst_name state_abbr
1 100654           Alabama A & M University         AL
2 100663 University of Alabama at Birmingham         AL
3 100690                 Amridge University         AL
```

```
4 100706 University of Alabama in Huntsville         AL
5 100724            Alabama State University         AL
6 100751           The University of Alabama         AL
  pred_degree_awarded_ipeds year earnings_med count_not_working count_working
1                         3 2007        36600               116          1139
2                         3 2007        40800               366          2636
3                         3 2007           NA                 6            25
4                         3 2007        49300               122           975
5                         3 2007        30500               210          1577
6                         3 2007        46700               292          2754
```

```
# cleaning
scorecard$region <- state.region[match(scorecard$state_abbr, state.abb)]
scorecard <- na.omit(scorecard)

head(scorecard)
```

```
  unitid                           inst_name state_abbr
1 100654            Alabama A & M University         AL
2 100663 University of Alabama at Birmingham         AL
4 100706 University of Alabama in Huntsville         AL
5 100724            Alabama State University         AL
6 100751           The University of Alabama         AL
7 100760   Central Alabama Community College         AL
  pred_degree_awarded_ipeds year earnings_med count_not_working count_working
1                         3 2007        36600               116          1139
2                         3 2007        40800               366          2636
4                         3 2007        49300               122           975
5                         3 2007        30500               210          1577
6                         3 2007        46700               292          2754
7                         2 2007        28100               113           590
  region
1  South
2  South
4  South
5  South
```

```
6   South
7   South
```

Research Questions

1. Does median income have a positive relationship with the proportion of working graduates?

H: Median income will have a positive relationship with the number of working graduates.

A positive relationship and significant p-value will prove this to be true.

2. Which US region contributes most to median earnings?

H0: All regions do not differ significantly for median earnings

HA: Eastern region will be the most significant in median earnings compared to other regions.

3. Which degree length leads to higher median salary?

H0: Median salary does not significantly differ between degree lengths.

HA: People with 4 year degrees have higher median salaries compared to other degre

Data Exploration

Manipulate Data

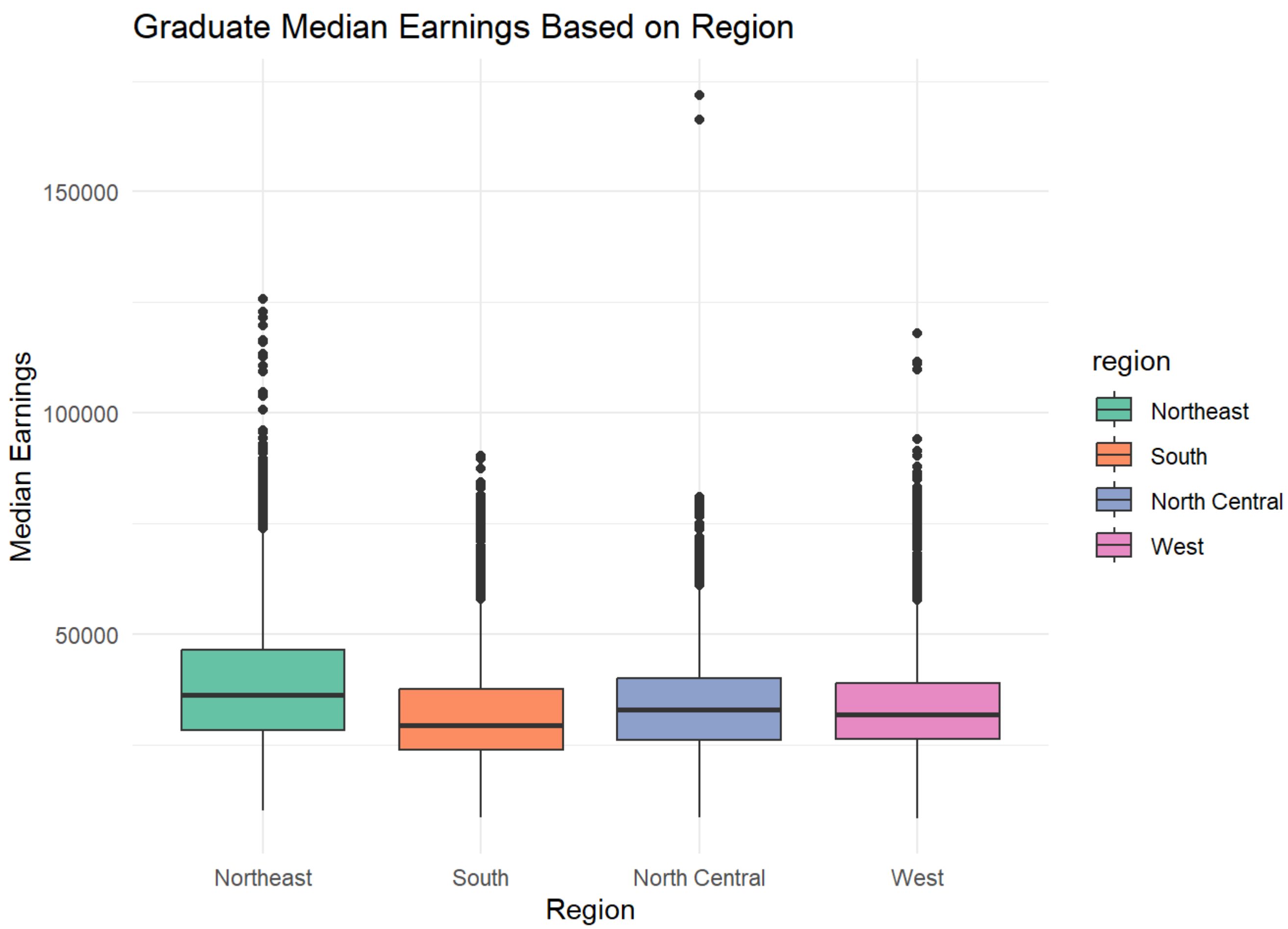
```
# Organize states into regions
scorecard$region <- state.region[match(scorecard$state_abbr, state.abb)]
scorecard <- na.omit(scorecard)

# Change variable name to 'degree'
scorecard = scorecard %>%
  mutate(degree = as.factor(pred_degree_awarded_ipeds))
```

```r
glimpse(scorecard)
```

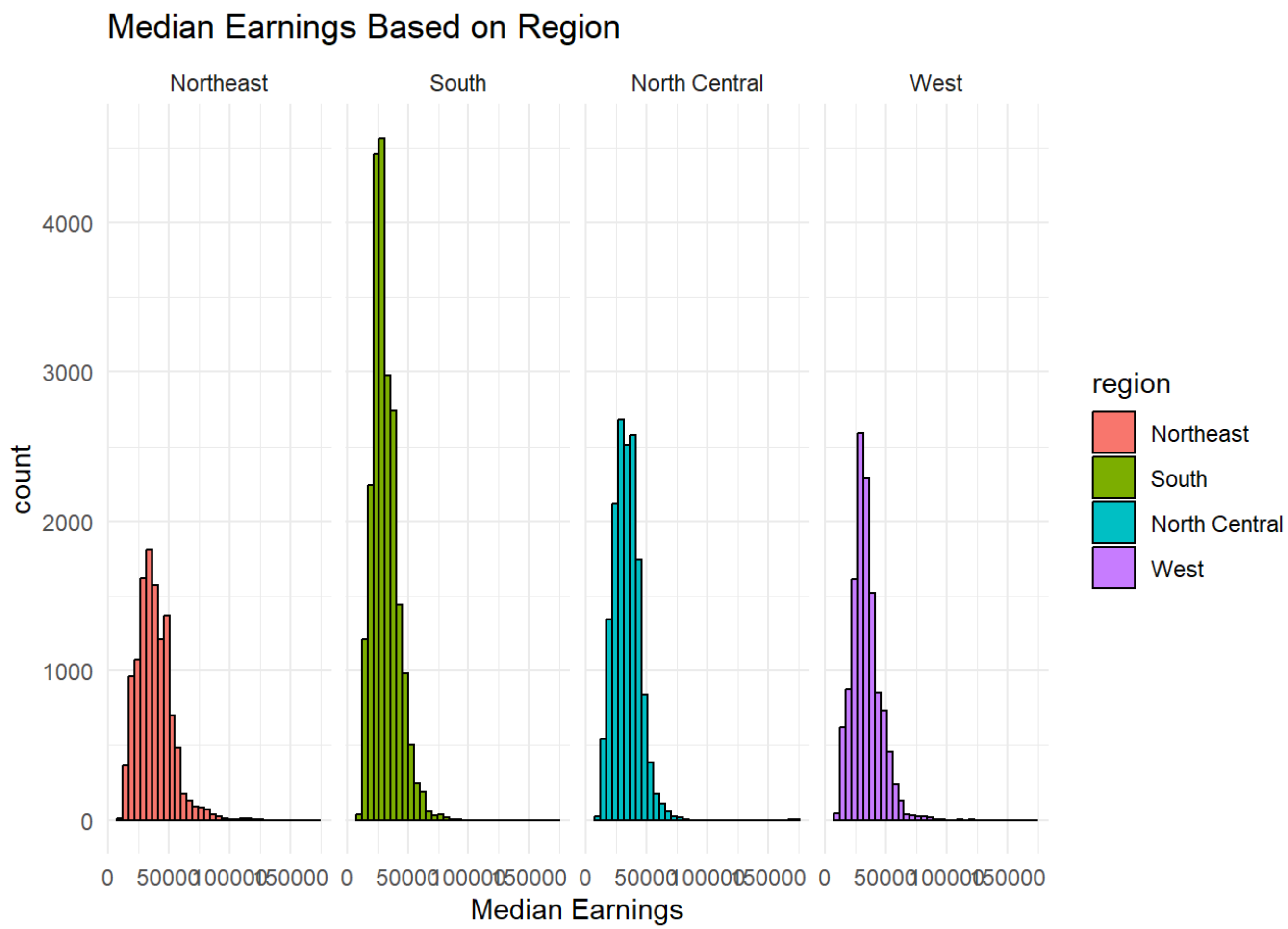
```
Rows: 30,401
Columns: 10
$ unitid                    <int> 100654, 100663, 100706, 100724, 100751, 1007…
$ inst_name                 <chr> "Alabama A & M University", "University of A…
$ state_abbr                <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "A…
$ pred_degree_awarded_ipeds <int> 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 2, 3, 3, 2, 2,…
$ year                      <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 20…
$ earnings_med              <int> 36600, 40800, 49300, 30500, 46700, 28100, 41…
$ count_not_working         <int> 116, 366, 122, 210, 292, 113, 77, 193, 348, …
$ count_working             <int> 1139, 2636, 975, 1577, 2754, 590, 676, 1400,…
$ region                    <fct> South, South, South, South, South, South, So…
$ degree                    <fct> 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 2, 3, 3, 2, 2,…
```

Region Exploration

```r
scorecard %>% ggplot(aes(x = region, y = earnings_med, fill = region)) +
  geom_boxplot() +
  labs(title = "Graduate Median Earnings Based on Region",
       x = "Region",
       y = "Median Earnings" ) +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()
```




```
scorecard1 <- scorecard %>% gather(key = "employment_status", value = "working", -c(unitid,inst_name,state

scorecard1 <- scorecard1 %>% mutate(employment_status = as.factor(employment_status))

scorecard1 %>%
  ggplot(aes(x = earnings_med, fill = region)) +
  geom_histogram(bins = 35, col = "black") + facet_grid(~region) +
  labs(title = "Median Earnings Based on Region",
       x = "Median Earnings") + facet_grid(~region) + theme_minimal()
```

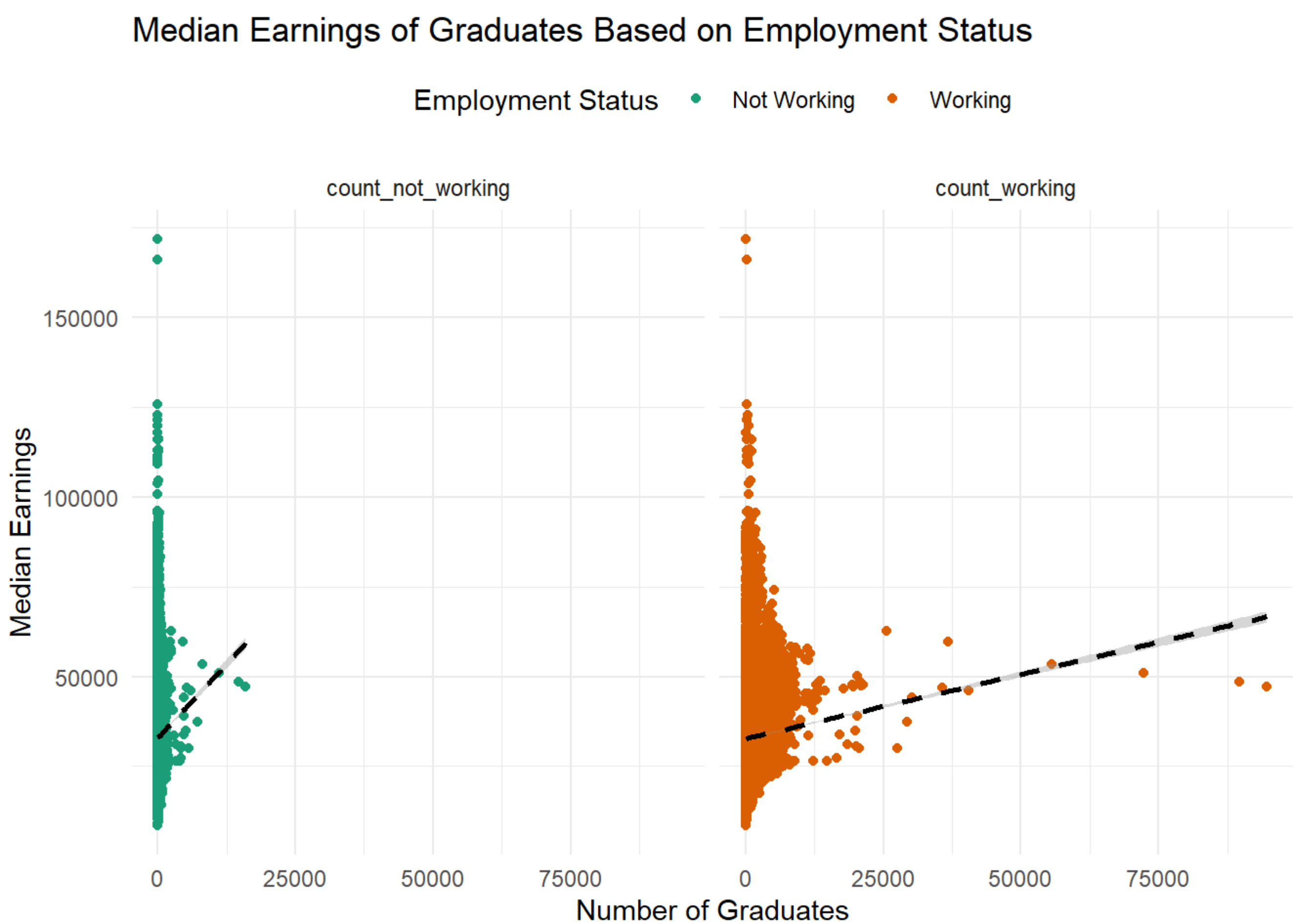


Employment Status vs. Median Earnings

```
# Create a new dataframe with tidy long performed on status of work
scorecard1 <- scorecard %>% gather(key = "employment_status", value = "working", -c(unitid,inst_name,state

scorecard1 <- scorecard1 %>% mutate(employment_status = as.factor(employment_status))

scorecard1 %>%
  ggplot(aes(x = working, y = earnings_med, color = employment_status)) +
  geom_point() +   facet_grid(~employment_status) +
  labs(title = "Median Earnings of Graduates Based on Employment Status",
       x = "Number of Graduates",
```

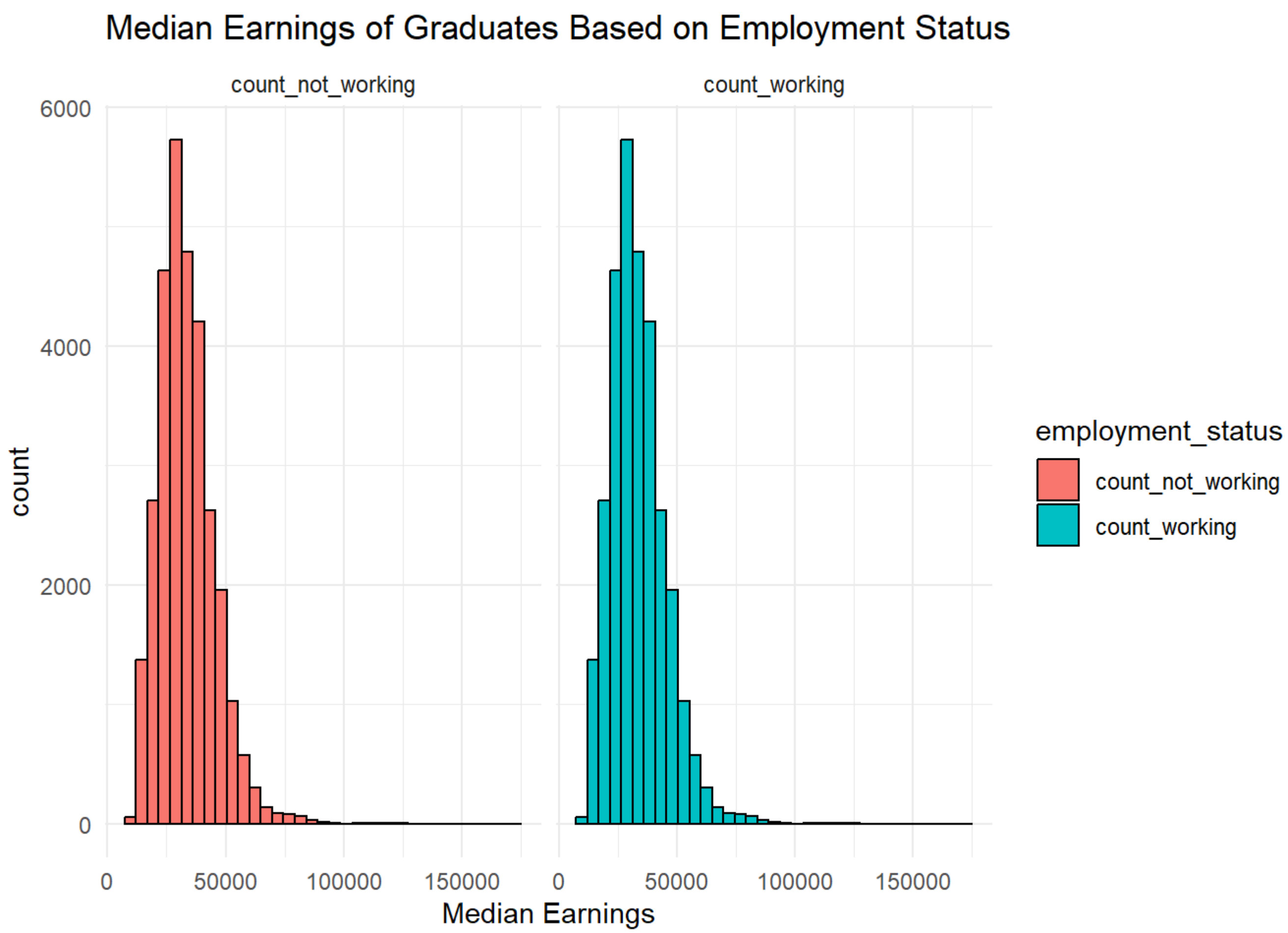
```
    y = "Median Earnings") +
  scale_color_brewer(palette = "Dark2", name = "Employment Status", labels = c("Not Working", "Working"))
  theme_minimal() +
  theme(legend.position = "top") +
  geom_smooth(method = "lm", linetype = "dashed", color = "black")
```

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




```
scorecard1 %>%
  ggplot(aes(x =  earnings_med, fill = employment_status)) +
```

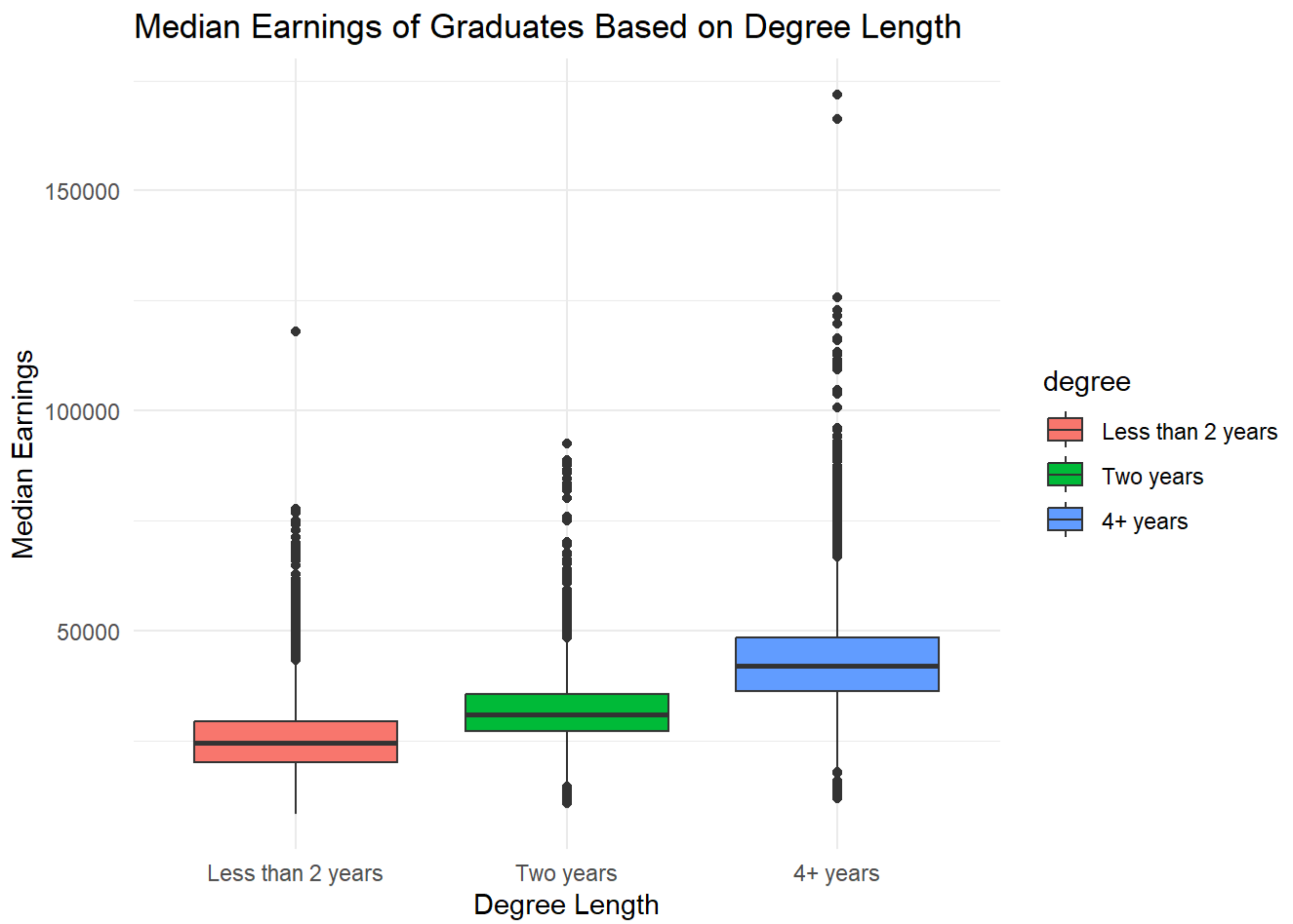
```r
  geom_histogram(bins = 35, col = "black") + facet_grid(~employment_status) +
  labs(title = "Median Earnings of Graduates Based on Employment Status",
       x = "Median Earnings") +
  scale_color_brewer(palette = "Dark2", name = "Employment Status", labels = c("Not Working", "Working"))
  theme_minimal()
```



Degree Length vs. Median Earnings

```r
scorecard1 %>%
  ggplot(aes(x = degree, y = earnings_med, fill = degree)) +
  geom_boxplot() +
```

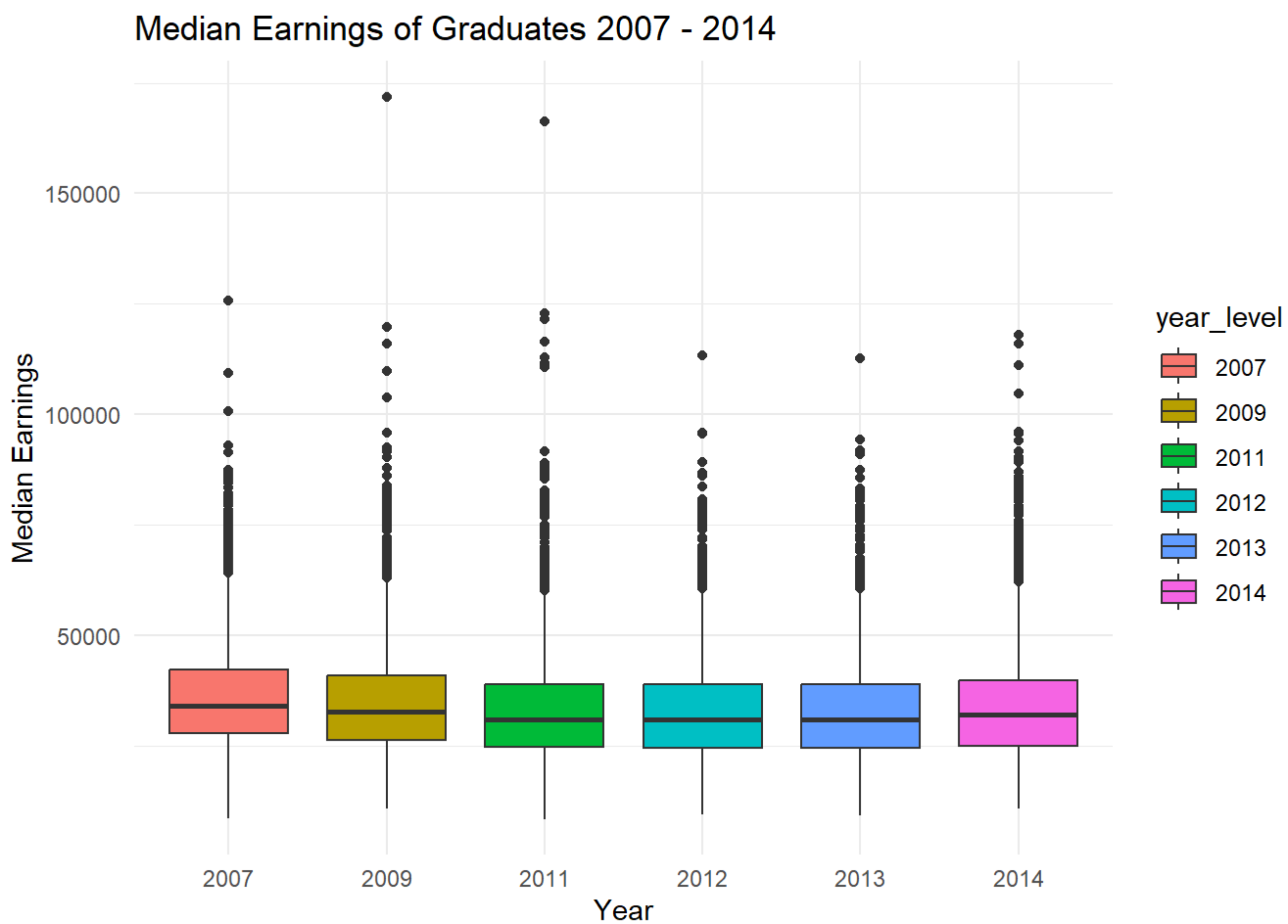
```
labs(title = "Median Earnings of Graduates Based on Degree Length",
     x = "Degree Length", y = "Median Earnings") +  theme_minimal() + scale_fill_discrete(labels=c('Less
scale_x_discrete(labels = c('Less than 2 years', 'Two years', '4+ years'))
```



Year vs. Median Earnings

```
scorecard1$year_level <- as.factor(scorecard1$year)
scorecard1 %>%
  ggplot(aes(x = year_level, y = earnings_med, fill = year_level)) +
  geom_boxplot() +
```

```r
  labs(title = "Median Earnings of Graduates 2007 - 2014",
       x = "Year", y = "Median Earnings") +  theme_minimal()
```

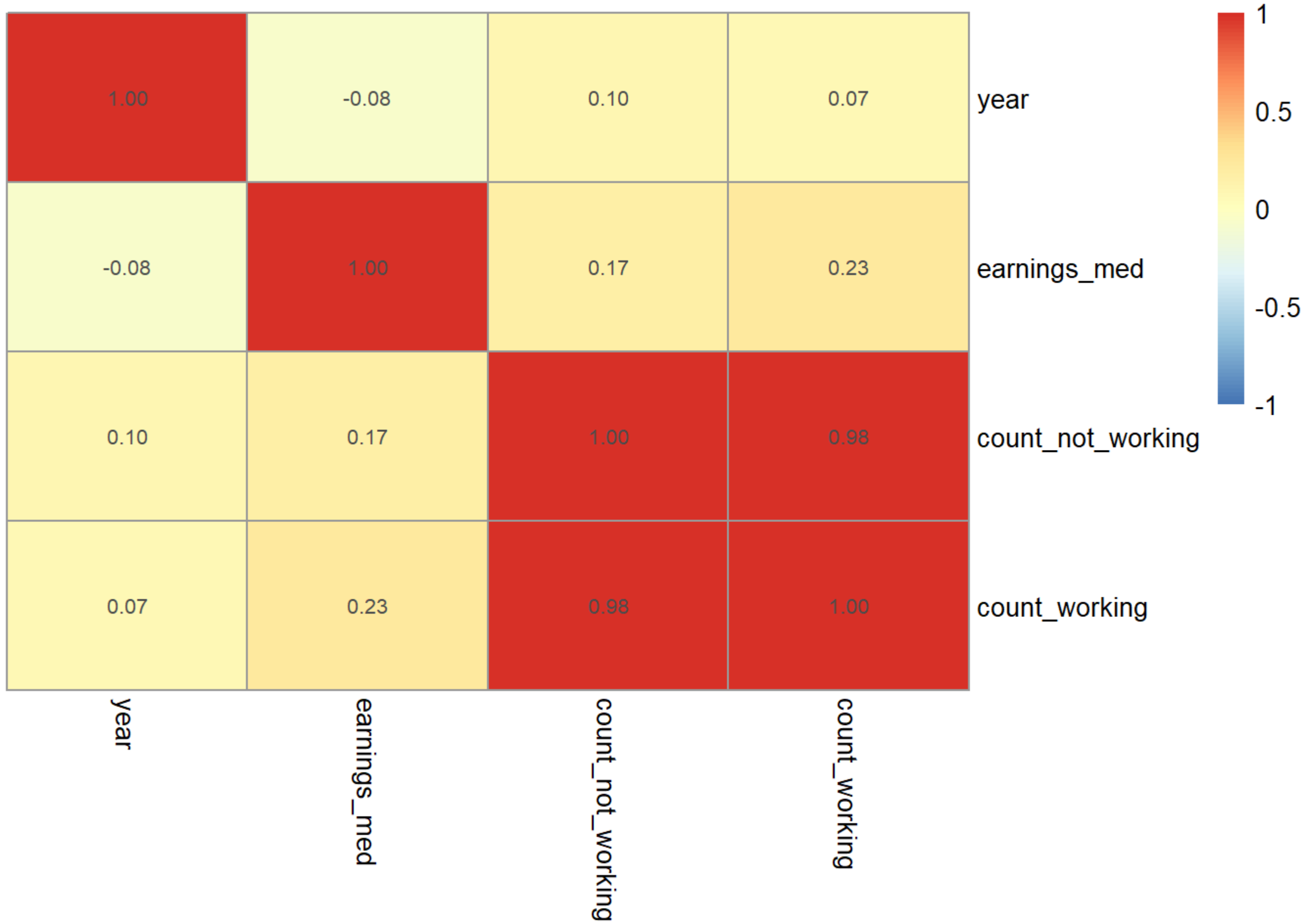


Correlations

```r
scorecard$year_num <- as.numeric(scorecard$year)
library(pheatmap)
```

```
Warning: package 'pheatmap' was built under R version 4.3.2
```

```r
scorecard_cor <- cor(na.omit(scorecard[,c(5, 6, 7, 8)]))
pheatmap(scorecard_cor,
        treeheight_col = 0,
        treeheight_row = 0,
        display_numbers = TRUE,
        breaks = seq(-1, 1, length = 101))
```



Multiple Linear Regression Model

To assess for the presence of a predictive relationship between the median earnings of individuals graduating from colleges and universities across the United States and characteristics associated with their alma mater and post college lives, we constructed a linear model regressing median earnings on surveyed universities' regional location, the number of alumni both employed and not working (not necessarily un-employed), the primary degree awarded, and the year that each survey was conducted.

```r
earnings_lm<-lm(earnings_med~region+degree+year+count_not_working+count_working, data=scorecard)
summary(earnings_lm)
```

```
Call:
lm(formula = earnings_med ~ region + degree + year + count_not_working + 
    count_working, data = scorecard)

Residuals:
   Min     1Q Median     3Q    Max 
-33245  -5048   -660   3946 130337 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          5.034e+05  4.173e+04   12.06   <2e-16 ***
regionSouth         -4.613e+03  1.364e+02  -33.82   <2e-16 ***
regionNorth Central -3.651e+03  1.456e+02  -25.07   <2e-16 ***
regionWest          -1.761e+03  1.552e+02  -11.35   <2e-16 ***
degree2              5.892e+03  1.220e+02   48.30   <2e-16 ***
degree3              1.567e+04  1.198e+02  130.82   <2e-16 ***
year                -2.359e+02  2.075e+01  -11.37   <2e-16 ***
count_not_working   -8.859e+00  2.397e-01  -36.97   <2e-16 ***
count_working        1.555e+00  3.766e-02   41.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8373 on 30392 degrees of freedom
Multiple R-squared:  0.4928,    Adjusted R-squared:  0.4927 
F-statistic:  3691 on 8 and 30392 DF,  p-value: < 2.2e-16
```

```
scorecard$region<-relevel(scorecard$region, ref="Northeast")
cat("Earnings median range:", range(scorecard$earnings_med))
```

```
Earnings median range: 8400 171900
```

The substantial F-statistic generated by the linear model of 3691 on 8 and 30392 degrees of freedom allowed us to reject the null hypothesis that none of the chosen variables possess any relationship to median earnings (all slopes are equal to zero) in favor of the alternative hypothesis that at least one of the predictive variables influences the earnings of American college graduates (at least one slope is not equal to zero). Given the confirmation of, at minimum, one of our independent variables' predictive power, we further explored the more nuanced ways in which each contributed to variation from the baseline predicted income of $503,340, as denoted by the intercept regression coefficient. Holding the influence of region, degrees typically awarded, year, and the number of graduates not actively employed constant, a one person increase in the number of gainfully employed graduates contributed to an institution results in a marginal $1.56 increase in predicted median earnings. Conversely, when controlling for the effect of all other predictors, the addition of a single non-working alumni unsurprisingly elicits a predicted $8.86 decline in predicted income. Assessment of the regression coefficient assigned to the year variable in the same manner revealed a slightly more impactful association between the year participants were surveyed and median earnings, with the passage of one year resulting in a loss of $235.90. Due to the categorical nature of the predominant degree awarded by collegiate study participants and the region in which each institution of higher learning resides, the analysis of their influence on predicted monetary outcomes diverged from that of aforementioned variables. As a hub for a variety of prestigious Universities, we anticipated that graduates from Northeastern schools would likely possess the highest median earnings and we accordingly designated it as the reference for our analysis of regional impacts. When controlling for the effects of all other variables and regions, prior attendance of a Southern school resulted in an average median earnings reduction of $4,613 from the Northeastern baseline. Upon similar evaluation, graduation from North Central and Western colleges comparably resulted in an average loss of $3,651 and $1,761, respectively. In considering the impact of the predominant degree awarded we identified the widest range of variation between predicted monetary outcomes, with the reference of less than 2 years differing by ampler amounts than the deviations observed between the regional categories. Controlling for all other variables and education levels, completion of a 2 year degree improved average predicted median income by $5,892, while graduation with a bachelor's degree raised income by an average of $15,670 after comparison to the baseline. Though all of the regression coefficients for both numeric and categorical variables possessed p-values significant at the zero level (p <2*10-16), the multiple R2 value of 0.498 indicates that only approximately 50% of the variation observed in median earnings for those surveyed is accounted for by the collegiate attributes analyzed above. This is reflected by the substantial residual standard error of 8373 on 30,392 degrees of freedom, meaning that the predicted values produced by the linear model deviate from actual monetary outcomes by an average of $8373. When compared to both the regression coefficients and the overall range of the actual median earnings values ($8604-$171900), the level of error observed in the estimates produced by the model is concerning and likely indicative of improper model fit through overfitting or multicollinearity.

Improving the Model

(Your text here)

Formal Hypothesis Tests

At the start of the paper we wanted to investigate how the number of working graduates (`count_working`), region of the university (`region`), and degree type (`degree`) relate to the median income of graduates earnings_med. In this section, we tested if each of these three variables are significant in predicting the median income. Firstly, we used the following equation to represent the relationship between median earnings and the chosen predictors:

$$Y = \beta_0 + \beta_{r_1}X_{r_1} + \beta_{r_2}X_{r_2} + \beta_{r_3}X_{r_3} + \beta_{d_1}X_{d_2} + \beta_{d_2}X_{d_2} + \beta_y X_y + \beta_n X_n + \beta_w X_w + \epsilon$$

Where: Y = `earnings_med` , X_r = `region`, X_d = `degree`, X_y = `year`, X_n = `count_not_working`, and X_w = `count_working` Using our final model, `earnings_lm`, we performed the following hypotheses testing:

For `region`:

- H_0: $\beta_{r_1} = \beta_{r_2} = \beta_{r_3} = 0$
- H_a: $\beta_{r1} \neq \beta_{r2} \neq \beta_{r3} \neq 0$

For `degree`:

- H_0: $\beta_{d_1} = \beta_{d_2} = 0$
- H_a: $\beta_{d_1} \neq \beta_{d_2} \neq 0$

For `count_working`:

- H_0: $\beta_w = 0$
- H_a: $\beta_w \neq 0$

Using the p-values from the `drop1` function, we see that β_r, β_g, and β_w are all significant predictors of `earnings_med`.

```
drop1(earnings_lm, test = "F")
```

```
Single term deletions

Model:
earnings_med ~ region + degree + year + count_not_working + count_working
                  Df  Sum of Sq        RSS    AIC F value    Pr(>F)    
<none>                          2.1307e+12 549219                      
region             3 9.3190e+10 2.2238e+12 550514  443.09 < 2.2e-16 ***
degree             2 1.2069e+12 3.3375e+12 562859 8607.45 < 2.2e-16 ***
year               1 9.0561e+09 2.1397e+12 549346  129.18 < 2.2e-16 ***
count_not_working  1 9.5798e+10 2.2265e+12 550554 1366.48 < 2.2e-16 ***
count_working      1 1.1952e+11 2.2502e+12 550876 1704.88 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Furthermore, using the `summary` functions we see that β_{r_1}, β_{r_2}, β_{r_3}, β_{d_1}, and β_{d_2} are all significant predictors of Y. We therefore reject H_0 for both `region` and `degree` and conclude that median income changes based on the regional location of the college and the type of degree the college offers. Also, we see that there is significant evidence that β_w is positive (which confirms our hypothesis in Part 1). We therefore reject H_0 for `count_working` variable and conclude that median earnings tend to increase as the number of working graduates increases.

```
summary(earnings_lm)
```

```

Call:
lm(formula = earnings_med ~ region + degree + year + count_not_working + 
    count_working, data = scorecard)

Residuals:
   Min     1Q Median     3Q    Max 
```

```
-33245  -5048   -660   3946 130337

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          5.034e+05  4.173e+04   12.06   <2e-16 ***
regionSouth         -4.613e+03  1.364e+02  -33.82   <2e-16 ***
regionNorth Central -3.651e+03  1.456e+02  -25.07   <2e-16 ***
regionWest          -1.761e+03  1.552e+02  -11.35   <2e-16 ***
degree2              5.892e+03  1.220e+02   48.30   <2e-16 ***
degree3              1.567e+04  1.198e+02  130.82   <2e-16 ***
year                -2.359e+02  2.075e+01  -11.37   <2e-16 ***
count_not_working   -8.859e+00  2.397e-01  -36.97   <2e-16 ***
count_working        1.555e+00  3.766e-02   41.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8373 on 30392 degrees of freedom
Multiple R-squared:  0.4928,    Adjusted R-squared:  0.4927 
F-statistic:  3691 on 8 and 30392 DF,  p-value: < 2.2e-16
```

To further investigate the `region` and `degree` variables, we ran the respective `constrast` functions and found that income significantly varies between all regions as well as between all degree types which confirms our initial hypothesis stated in Part 1.

```
cat("Comparing median income between regions:", "\n")
```

```
Comparing median income between regions:
```

```
contrast(emmeans(earnings_lm, ~ region), method = "pairwise", adjust = "none")
```

```
 contrast                  estimate  SE    df t.ratio p.value
 Northeast - South             4613 136 30392  33.816  <.0001
 Northeast - North Central     3651 146 30392  25.072  <.0001
 Northeast - West              1761 155 30392  11.348  <.0001
 South - North Central         -962 126 30392  -7.644  <.0001
 South - West                 -2852 135 30392 -21.172  <.0001
 North Central - West         -1890 145 30392 -12.998  <.0001
```

```
Results are averaged over the levels of: degree
```

```r
cat("\n","Comparing median income between degrees:", "\n", sep = "")
```

```
Comparing median income between degrees:
```

```r
contrast(emmeans(earnings_lm, ~ degree), method = "pairwise", adjust = "none")
```

```
 contrast          estimate  SE    df  t.ratio p.value
 degree1 - degree2    -5892 122 30392  -48.304  <.0001
 degree1 - degree3   -15673 120 30392 -130.823  <.0001
 degree2 - degree3    -9781 128 30392  -76.311  <.0001

Results are averaged over the levels of: region
```

In conclusion, based on our findings, all of our initial hypotheses seem to be confirmed. The median earnings do seem to increase with the number of graduates that are able to find a job. The earnings also vary based on degree type the graduate received and the geographic region of the US where the college is located. These conclusions do have serious limitations though. Firstly, our model contained only 5 predictors all of which were found to be significant. However, the inclusion of more predictors can affect the trends of the model and change the significance of each of the original 5 predictors. Also, we need to consider the possibility of existence of confounding variables. For example, it is possible that graduates who go to elite colleges are more likely to both find a job and earn a higher wage. Also, some regions in the US like the Northeast tend to have many states with a significantly higher cost of living which can explain the difference in median earnings. It is also important to account for the fact that we performed multiple tests in this section, hence we adjusted our p-values using the Bonferroni correction. Firstly, for both region and degree variables, we repeated the pairwise comparisons but using the Bonferroni adjusted p-values. In both cases, our conclusions did not change.

```r
cat("Comparing median income between regions:", "\n")
```

```
Comparing median income between regions:
```

```r
contrast(emmeans(earnings_lm, ~ degree), method = "pairwise", adjust = "bonferroni")
```

```
 contrast          estimate  SE    df  t.ratio p.value
 degree1 - degree2    -5892 122 30392  -48.304  <.0001
 degree1 - degree3   -15673 120 30392 -130.823  <.0001
 degree2 - degree3    -9781 128 30392  -76.311  <.0001

Results are averaged over the levels of: region
P value adjustment: bonferroni method for 3 tests
```

```r
cat("\n", "Comparing median income between degrees:", "\n", sep="")
```

```
Comparing median income between degrees:
```

```r
contrast(emmeans(earnings_lm, ~ region), method = "pairwise", adjust = "bonferroni")
```

```
 contrast                  estimate  SE    df t.ratio p.value
 Northeast - South             4613 136 30392  33.816  <.0001
 Northeast - North Central     3651 146 30392  25.072  <.0001
 Northeast - West              1761 155 30392  11.348  <.0001
 South - North Central         -962 126 30392  -7.644  <.0001
 South - West                 -2852 135 30392 -21.172  <.0001
 North Central - West         -1890 145 30392 -12.998  <.0001

Results are averaged over the levels of: degree
P value adjustment: bonferroni method for 6 tests
```

Then, since we tested three separate sets of hypotheses, the resulting p-values had to be multiplied by a factor of 3 to perform the Bonferroni correction. However, in all three cases we ended up with a $p - value < 2 * 10^{-16}$ so it follows that we still must reject H_0 in all three cases.

Robustness of Results

(Your text here)

Conclusions

(Your text here)