Analysis on College Major selection depending on Income

Introduction

In this analysis we will analyse if there is a correlation between income and the selection of college major categories.

A codebook for the dataset is given below:

- rank: Rank by median earnings
- major_code: Major code
- major: Major description
- major_category: Category of major
- total: Total number of people with major
- sample_size: Sample size of full-time, year-round individuals used for income/earnings estimates: p25th, median, p75th
- p25th: 25th percentile of earnings
- median: Median earnings of full-time, year-round workers
- p75th: 75th percentile of earnings
- perc_men: % men with major (out of total)
- perc_women: % women with major (out of total)
- perc_employed: % employed (out of total)
- perc_employed_fulltime: % employed 35 hours or more (out of employed)
- perc_employed_parttime: % employed less than 35 hours (out of employed)
- perc_employed_fulltime_yearround: % employed at least 50 weeks and at least 35 hours (out of employed and full-time)
- perc_unemployed: % unemployed (out of employed)
- perc_college_jobs: % with job requiring a college degree (out of employed)
- perc_non_college_jobs: % with job not requiring a college degree (out of employed)
- perc_low_wage_jobs: % in low-wage service jobs (out of total)

Library loading and Data reading

```
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
```

```
library(ggplot2)
library(collegeIncome)
data(college)
```

Exploratory data analysis

```
head(college)
```

```
##   rank major_code                                    major major_category
## 1    1       2419                      Petroleum Engineering    Engineering
## 2    2       2416             Mining And Mineral Engineering    Engineering
## 3    3       2415                  Metallurgical Engineering    Engineering
## 4    4       2417 Naval Architecture And Marine Engineering    Engineering
## 5    5       2405                       Chemical Engineering    Engineering
## 6    6       2418                        Nuclear Engineering    Engineering
##   total sample_size perc_women p25th median  p75th   perc_men perc_employed
## 1  2339          36  0.9109326 25000  40000  50000 0.08906743     0.9115044
## 2   756           7  0.5154064 26000  37000  40000 0.48459355     0.7980501
## 3   856           3  0.5942076 26700  45000  60000 0.40579235     0.7871943
## 4  1258          16  0.6521298 26000  35000  45000 0.34787018     0.8465608
## 5 32260         289  0.4179248 31500  62000 109000 0.58207520     0.8515625
## 6  2573          17  0.4305368 23000  44700  50000 0.56946324     0.8474507
##   perc_employed_fulltime perc_employed_parttime
## 1              0.9206524              0.1774785
## 2              0.7110092              0.3623853
## 3              0.8833498              0.3387257
## 4              0.9366337              0.1673267
## 5              0.8086363              0.4020061
## 6              0.8756262              0.2040405
##   perc_employed_fulltime_yearround perc_unemployed perc_college_jobs
## 1                        0.7704431      0.08849558         0.6702970
## 2                        0.7093101      0.20194986         0.3867764
## 3                        0.7738366      0.21280567         0.7289116
## 4                        0.6527853      0.15343915         0.2460902
## 5                        0.6852821      0.14843750         0.5867515
## 6                        0.6567727      0.15254929         0.4624782
##   perc_non_college_jobs perc_low_wage_jobs
## 1             0.1821782         0.05544554
## 2             0.5158761         0.21560172
## 3             0.1759983         0.03014828
## 4             0.4107636         0.04323827
## 5             0.3860437         0.11801062
## 6             0.4057592         0.23472949
```

```
str(college)
```

```
## 'data.frame':    173 obs. of  19 variables:
##  $ rank                          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ major_code                    : int  2419 2416 2415 2417 2405 2418 6202 5001 2414 2408 ...
##  $ major                         : chr  "Petroleum Engineering" "Mining And Mineral Engineering" "
##  $ major_category                : chr  "Engineering" "Engineering" "Engineering" "Engineering" ..
##  $ total                         : int  2339 756 856 1258 32260 2573 3777 1792 91227 81527 ...
##  $ sample_size                   : int  36 7 3 16 289 17 51 10 1029 631 ...
##  $ perc_women                    : num  0.911 0.515 0.594 0.652 0.418 ...
##  $ p25th                         : num  25000 26000 26700 26000 31500 23000 32500 37900 29200 2300
##  $ median                        : num  40000 37000 45000 35000 62000 44700 45000 57000 36000 3220
##  $ p75th                         : num  50000 40000 60000 45000 109000 50000 58000 67000 46000 471
##  $ perc_men                      : num  0.0891 0.4846 0.4058 0.3479 0.5821 ...
##  $ perc_employed                 : num  0.912 0.798 0.787 0.847 0.852 ...
```

```
##  $ perc_employed_fulltime          : num  0.921 0.711 0.883 0.937 0.809 ...
##  $ perc_employed_parttime          : num  0.177 0.362 0.339 0.167 0.402 ...
##  $ perc_employed_fulltime_yearround: num  0.77 0.709 0.774 0.653 0.685 ...
##  $ perc_unemployed                 : num  0.0885 0.2019 0.2128 0.1534 0.1484 ...
##  $ perc_college_jobs               : num  0.67 0.387 0.729 0.246 0.587 ...
##  $ perc_non_college_jobs           : num  0.182 0.516 0.176 0.411 0.386 ...
##  $ perc_low_wage_jobs              : num  0.0554 0.2156 0.0301 0.0432 0.118 ...
```

We can see that this dataframe contains 173 observations of 19 variables corresponding to the codebook. For my analysis only some of these variables are important including "major_category" and "median".

I will start by converting the "major_category" to a factor variable:

```
college$major_category <- as.factor(college$major_category)
```

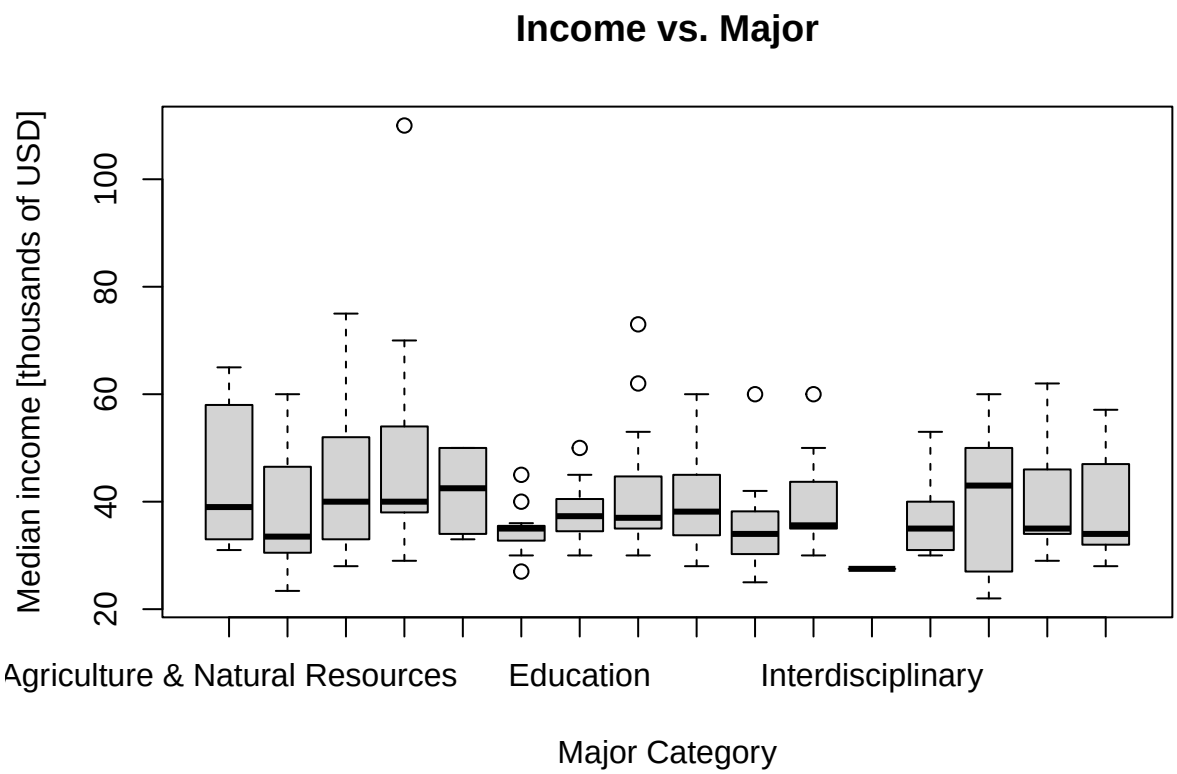
We can then see the unique values of the categories:

```
unique(college$major_category)
```

```
##  [1] Engineering                          Business
##  [3] Physical Sciences                    Law & Public Policy
##  [5] Computers & Mathematics              Agriculture & Natural Resources
##  [7] Industrial Arts & Consumer Services  Arts
##  [9] Health                               Social Science
## [11] Biology & Life Science               Education
## [13] Humanities & Liberal Arts            Psychology & Social Work
## [15] Communications & Journalism          Interdisciplinary
## 16 Levels: Agriculture & Natural Resources Arts ... Social Science
```

I then proceed to analyze the medians of the incomes by major category:

```
boxplot(median/1000 ~ major_category, data = college, main = "Income vs. Major", ylab = "Median income
```



We can induce from the boxplot that the distribution is skewed and not normal.

Statistical Data Analysis & Regression Model

We proceed to order the categories alphabetically and fit a linear model to compare each median with the first one (Agriculture & Natural Resources):

```
fit <- lm(median ~ major_category,data = college)
summary(fit)$coef
```

```
##                                                   Estimate Std. Error
## (Intercept)                                    43500.0000   3590.819
## major_categoryArts                             -5450.0000   5386.228
## major_categoryBiology & Life Science             364.2857   4701.486
## major_categoryBusiness                          5653.8462   4776.236
## major_categoryCommunications & Journalism      -1500.0000   6717.807
## major_categoryComputers & Mathematics          -8781.8182   4961.429
## major_categoryEducation                        -5562.5000   4577.414
## major_categoryEngineering                      -3106.8966   4164.154
## major_categoryHealth                           -3183.3333   4861.992
## major_categoryHumanities & Liberal Arts        -8333.3333   4635.727
## major_categoryIndustrial Arts & Consumer Services  -3071.4286   5595.887
## major_categoryInterdisciplinary               -16000.0000  11909.399
## major_categoryLaw & Public Policy              -5700.0000   6219.481
## major_categoryPhysical Sciences                -3100.0000   5078.185
## major_categoryPsychology & Social Work         -3611.1111   5217.339
## major_categorySocial Science                   -4433.3333   5217.339
##                                                    t value     Pr(>|t|)
## (Intercept)                                    12.11422804 2.873928e-24
## major_categoryArts                             -1.01183974 3.131715e-01
## major_categoryBiology & Life Science            0.07748311 9.383379e-01
## major_categoryBusiness                          1.18374520 2.383031e-01
## major_categoryCommunications & Journalism      -0.22328715 8.236023e-01
## major_categoryComputers & Mathematics          -1.77001776 7.866520e-02
## major_categoryEducation                        -1.21520579 2.261119e-01
## major_categoryEngineering                      -0.74610504 4.567197e-01
## major_categoryHealth                           -0.65473851 5.135942e-01
## major_categoryHumanities & Liberal Arts        -1.79763232 7.415704e-02
## major_categoryIndustrial Arts & Consumer Services -0.54887249 5.838727e-01
## major_categoryInterdisciplinary                -1.34347667 1.810563e-01
## major_categoryLaw & Public Policy              -0.91647520 3.608233e-01
## major_categoryPhysical Sciences                -0.61045434 5.424435e-01
## major_categoryPsychology & Social Work         -0.69213657 4.898739e-01
## major_categorySocial Science                   -0.84973074 3.967687e-01
```

From these coefficients we can induce that the median income for Agriculture students is $43500, and that there is no significant difference between the income medians of all other major categories is not statistically significant (using a p-value of 0.05).

Since business has the highest median income, we relevel the fit to compare all other categories with it:

```
major_category_b <- relevel(college$major_category, "Business")
fitb <- lm(median ~ major_category_b, data = college)
summary(fitb)$coef
```

```
##                                                    Estimate Std. Error
## (Intercept)                                     49153.846   3149.357
## major_category_bAgriculture & Natural Resources -5653.846   4776.236
## major_category_bArts                           -11103.846   5102.541
```

```
## major_category_bBiology & Life Science                -5289.560   4373.606
## major_category_bCommunications & Journalism           -7153.846   6492.565
## major_category_bComputers & Mathematics              -14435.664   4651.908
## major_category_bEducation                            -11216.346   4239.951
## major_category_bEngineering                           -8760.743   3790.072
## major_category_bHealth                                -8837.179   4545.705
## major_category_bHumanities & Liberal Arts            -13987.179   4302.840
## major_category_bIndustrial Arts & Consumer Services   -8725.275   5323.384
## major_category_bInterdisciplinary                    -21653.846  11783.813
## major_category_bLaw & Public Policy                  -11353.846   5975.484
## major_category_bPhysical Sciences                     -8753.846   4776.236
## major_category_bPsychology & Social Work              -9264.957   4923.931
## major_category_bSocial Science                       -10087.179   4923.931
##                                                        t value     Pr(>|t|)
## (Intercept)                                          15.607584 9.444322e-34
## major_category_bAgriculture & Natural Resources      -1.183745 2.383031e-01
## major_category_bArts                                 -2.176141 3.103954e-02
## major_category_bBiology & Life Science               -1.209428 2.283166e-01
## major_category_bCommunications & Journalism          -1.101852 2.722123e-01
## major_category_bComputers & Mathematics              -3.103171 2.271210e-03
## major_category_bEducation                            -2.645395 8.989341e-03
## major_category_bEngineering                          -2.311498 2.210557e-02
## major_category_bHealth                               -1.944073 5.367450e-02
## major_category_bHumanities & Liberal Arts            -3.250685 1.408831e-03
## major_category_bIndustrial Arts & Consumer Services -1.639047 1.032059e-01
## major_category_bInterdisciplinary                    -1.837592 6.801278e-02
## major_category_bLaw & Public Policy                  -1.900071 5.925698e-02
## major_category_bPhysical Sciences                    -1.832792 6.872781e-02
## major_category_bPsychology & Social Work             -1.881618 6.173891e-02
## major_category_bSocial Science                       -2.048603 4.216615e-02
```

```
pval <- summary(fitb)$coef[,4] < 0.025
pval
```

```
##                                         (Intercept)
##                                                TRUE
##     major_category_bAgriculture & Natural Resources
##                                               FALSE
##                                  major_category_bArts
##                                               FALSE
##              major_category_bBiology & Life Science
##                                               FALSE
##         major_category_bCommunications & Journalism
##                                               FALSE
##             major_category_bComputers & Mathematics
##                                                TRUE
##                            major_category_bEducation
##                                                TRUE
##                          major_category_bEngineering
##                                                TRUE
##                               major_category_bHealth
##                                               FALSE
##           major_category_bHumanities & Liberal Arts
##                                                TRUE
## major_category_bIndustrial Arts & Consumer Services
```

```
##                                             FALSE
##                   major_category_bInterdisciplinary
##                                             FALSE
##                 major_category_bLaw & Public Policy
##                                             FALSE
##                   major_category_bPhysical Sciences
##                                             FALSE
##           major_category_bPsychology & Social Work
##                                             FALSE
##                      major_category_bSocial Science
##                                             FALSE
```

```
fit4l <- lm(median ~ major_category_b - 1, data = college)
summary(fit4l)$coef
```

```
##                                                 Estimate Std. Error
## major_category_bBusiness                        49153.85   3149.357
## major_category_bAgriculture & Natural Resources 43500.00   3590.819
## major_category_bArts                            38050.00   4014.658
## major_category_bBiology & Life Science          43864.29   3034.796
## major_category_bCommunications & Journalism     42000.00   5677.583
## major_category_bComputers & Mathematics         34718.18   3423.712
## major_category_bEducation                       37937.50   2838.792
## major_category_bEngineering                     40393.10   2108.602
## major_category_bHealth                          40316.67   3277.954
## major_category_bHumanities & Liberal Arts       35166.67   2931.891
## major_category_bIndustrial Arts & Consumer Services 40428.57   4291.850
## major_category_bInterdisciplinary               27500.00  11355.167
## major_category_bLaw & Public Policy             37800.00   5078.185
## major_category_bPhysical Sciences               40400.00   3590.819
## major_category_bPsychology & Social Work        39888.89   3785.056
## major_category_bSocial Science                  39066.67   3785.056
##                                                   t value     Pr(>|t|)
## major_category_bBusiness                        15.607584 9.444322e-34
## major_category_bAgriculture & Natural Resources 12.114228 2.873928e-24
## major_category_bArts                             9.477769 3.919976e-17
## major_category_bBiology & Life Science          14.453784 1.191360e-30
## major_category_bCommunications & Journalism      7.397514 7.850192e-12
## major_category_bComputers & Mathematics         10.140510 6.691567e-19
## major_category_bEducation                       13.363961 1.095127e-27
## major_category_bEngineering                     19.156348 6.199089e-43
## major_category_bHealth                          12.299338 8.947526e-25
## major_category_bHumanities & Liberal Arts       11.994532 6.110648e-24
## major_category_bIndustrial Arts & Consumer Services  9.419848 5.577563e-17
## major_category_bInterdisciplinary                2.421805 1.658291e-02
## major_category_bLaw & Public Policy              7.443604 6.067423e-12
## major_category_bPhysical Sciences               11.250915 6.574192e-22
## major_category_bPsychology & Social Work        10.538521 5.658611e-20
## major_category_bSocial Science                  10.321293 2.183346e-19
```

In this case, we use 0.025 as a p-value since we want to see if the median income is significantly smaller or higher than that of Business major students, some categories do cross this threshold, these and their median income are:

Major Category	Median Income [USD]
Computers & Mathematics	34718.80
Education	37937.50
Engineering	40393.10
Humanities & Liberal Arts	35166.67

Conclusion

From the data we can conclude that there isn't enough evidence to probe that there is a significant correlation between income and major category. The only exception to this are Business majors, which do show a statistically significant difference between the 4 lower earning majors.