Final Project

Skyler Hauser

2023-06-08

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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 errors
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(rcompanion)
library(AER)
```

```
## Loading required package: car
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:openintro':
##
##     densityPlot
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
##
## Loading required package: lmtest
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: sandwich
## Loading required package: survival
##
## Attaching package: 'survival'
##
## The following object is masked from 'package:openintro':
##
##     transplant
```

SOC 212 Final Project: Where do college graduates live?

Introduction

The admissions office at HWS (Hobart and William Smith Colleges) has tasked me with conducting research to understandwhere college graduates live as adults because they want to recruit in these areas. By analyzing data from the dataset called CollegeDistance, which consists of observations from college graduates who graduated from high school in 1986, I will aim to compare this data with their current data. The goal is to identify how patterns of social life have changed since 1986 and allocate resources more effectively to recruit in areas with a high potential for yielding new students.

In this report, I will conduct a series of tests and analyses to explore where college students from the 1986 cohort live as adults. I will be using these specific variables: Home, Does the family own their home?, measured in factor of yes and no. Unemp (unemployment), this is the countys unemployment rate in 1980, numerically measured. Distance, the distance from 4-year college (in 10 miles), numerically measured. Income, Is the family income above USD 25,000 per year? measured by factor of yes or no. And finally, wage, the state hourly wage in manufacturing in 1980, numerically measured.

For this report, I will conduct various tests and analyses to gain insights into the residential patterns of college graduates from the 1986 cohort. I will explore the relationships between these variables and evaluate their significance in understanding the geographic distribution of college graduates from that era.

While doing this statistical analysis, I will explore several research questions related to the dataset. First, I will examine the relationship between the level of income and homeownership among the 1986 college graduates. I will speculate on the potential association between income and homeownership and evaluate its existence in the broader population. Additionally, I will estimate the true population mean distance from a 4-year college for these respondents, as well as analyze the average difference in county unemployment rates based on income levels. I will also investigate the population proportion of women who own their own homes and discuss the hypothesized relationship between income, homeownership, and gender. Furthermore, I will explore the impact of a father's college attendance on the educational attainment of students and assess if there are significant differences between the respondents in this dataset and the population as a whole. Lastly, I will examine the relationship between state manufacturing wages and the distance respondents live from a 4-year institution, considering its potential implications for recruitment strategies at the admissions office.

The findings from this analysis will provide valuable information to the admissions office at HWS, allowing them to compare the current data with the older data and gain insights into how patterns of social life have changed since 1986. These insights will assist in better allocating resources for recruitment efforts in areas that have a high potential for yielding new students.

```
data("CollegeDistance")
?CollegeDistance
glimpse(CollegeDistance)
```

```
## Rows: 4,739
## Columns: 14
## $ gender    <fct> male, female, male, male, female, male, female, female, male~
## $ ethnicity <fct> other, other, other, afam, other, other, other, other, other~
## $ score     <dbl> 39.15, 48.87, 48.74, 40.40, 40.48, 54.71, 56.07, 54.85, 64.7~
## $ fcollege  <fct> yes, no, no, no, no, no, no, no, yes, no, no, no, no, yes, y~
## $ mcollege  <fct> no, no, no, no, no, no, no, no, no, no, no, yes, no, no, yes~
## $ home      <fct> yes, yes, yes, yes, no, yes, yes, yes, yes, yes, yes, yes, y~
## $ urban     <fct> yes, yes, yes, yes, yes, yes, no, no, yes, yes, yes, yes, no~
## $ unemp     <dbl> 6.2, 6.2, 6.2, 6.2, 5.6, 5.6, 7.2, 7.2, 5.9, 5.9, 5.9, 5.9, ~
## $ wage      <dbl> 8.09, 8.09, 8.09, 8.09, 8.09, 8.09, 8.85, 8.85, 8.09, 8.09, ~
## $ distance  <dbl> 0.2, 0.2, 0.2, 0.2, 0.4, 0.4, 0.4, 0.4, 3.0, 3.0, 3.0, 3.0, ~
## $ tuition   <dbl> 0.88915, 0.88915, 0.88915, 0.88915, 0.88915, 0.88915, 0.8498~
## $ education <dbl> 12, 12, 12, 12, 13, 12, 13, 15, 13, 15, 12, 14, 15, 17, 14, ~
## $ income    <fct> high, low, low, low, low, low, low, low, low, low, high, hig~
## $ region    <fct> other, other, other, other, other, other, other, other, othe~
```

Variables to Familiarize

```
table(CollegeDistance$home)
```

```
##
##   no  yes
##  852 3887
```

```
summary(CollegeDistance$unemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   5.900   7.100   7.597   8.900  24.900
```

```
summary(CollegeDistance$distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.400   1.000   1.803   2.500  20.000
```

```
table(CollegeDistance$income)
```

```
##
##  low high
## 3374 1365
```

```
summary(CollegeDistance$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.590   8.850   9.680   9.501  10.150  12.960
```

From this information, we learn the following about the four variables of interest for this assignments:

1. Most respondents family's have a home, 3,887 have a home and 852 respondents family's don't have a home.
2. The mean, average, country unemployment rate in 1980 is 7.60.
3. The median, middle of the set, for distance from 4-year college(goes by 10 miles) is 1.
4. Most respondents family income above 25,000 per year is low, which is 3374. The high is 1365.
5. The mean, average, wage, state hourly wage in manufacturing in 1980 is 9.680.

Variable: home

```
mode_home <- CollegeDistance%>%
  count(home)%>%
  arrange(desc(n))
mode_home
```

```
##   home    n
## 1  yes 3887
## 2   no  852
```

```
freqtable_home <- table(CollegeDistance$home)
freqtable_home
```

```
##
##   no  yes
##  852 3887
```

Measurement: This variable represents the home state of each student in the data set. Level of measurement: Nominal Variable. Summary: To summarize the distribution of home states, we can use a frequency table to count the number of students from each state. Since this is a Nominal variable, we can only calculate the mode.

Variable: unemp

```
summary_unemployment <- CollegeDistance%>%
  summarize(meanunemp = mean(unemp, na.rm = TRUE),
            medianunemp = median(unemp, na.rm = TRUE),
            sdunemp = sd(unemp, na.rm = TRUE),
            rangeunemp = range(unemp, na.rm = TRUE),
            iqrunemp = IQR(unemp, na.rm = TRUE))
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
summary_unemployment
```

```
##   meanunemp medianunemp  sdunemp rangeunemp iqrunemp
## 1  7.597215         7.1 2.763581        1.4        3
## 2  7.597215         7.1 2.763581       24.9        3
```

```
mode_unemp <- CollegeDistance%>%
  count(unemp)%>%
  arrange(desc(n))
mode_unemp
```

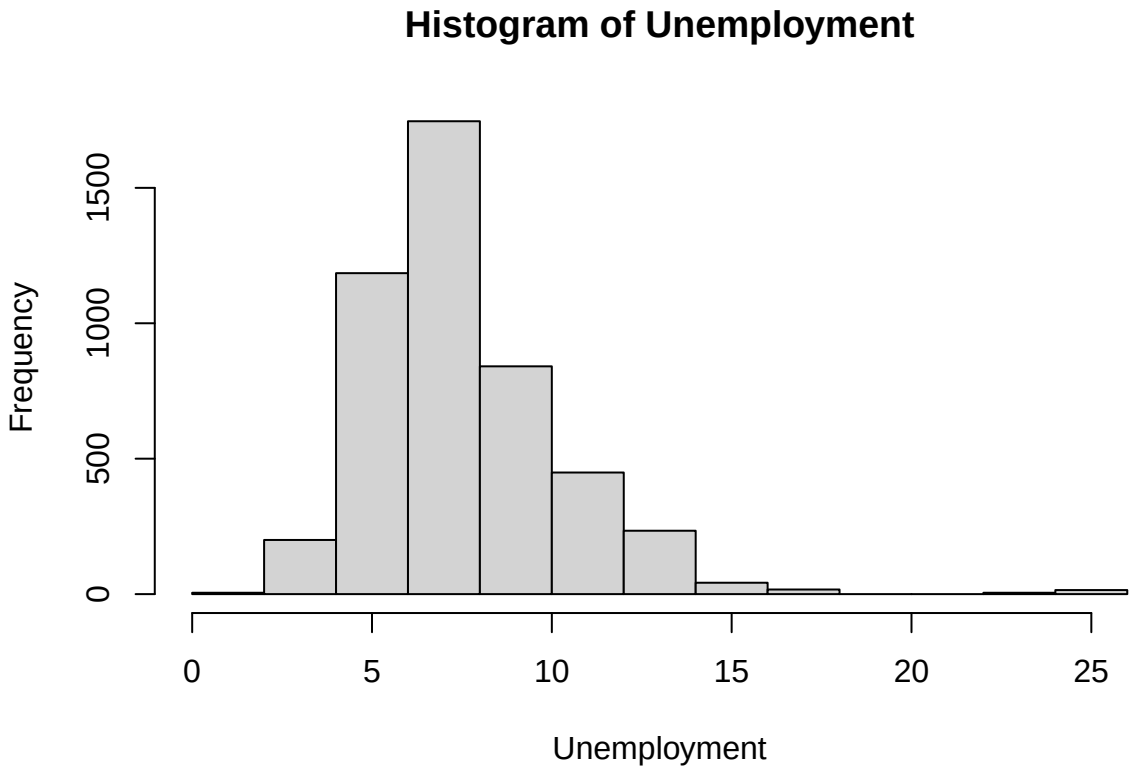
```
##     unemp   n
## 1     8.0 260
## 2     6.6 245
## 3     8.2 201
## 4     5.9 178
## 5     7.2 177
```

```
## 6     7.7 153
## 7     6.8 137
## 8     5.5 116
## 9     6.1 111
## 10    5.3 109
## 11    4.3 100
## 12    6.7  94
## 13    5.6  84
## 14    6.3  84
## 15    9.2  81
## 16    6.9  80
## 17    5.1  78
## 18    8.9  76
## 19    6.2  66
## 20    8.7  59
## 21    4.0  58
## 22    4.7  58
## 23    7.8  57
## 24   10.7  55
## 25    8.1  53
## 26    6.4  51
## 27    4.4  47
## 28    4.6  47
## 29    9.5  47
## 30   10.8  47
## 31   11.8  45
## 32    9.1  44
## 33    4.9  43
## 34    6.5  43
## 35   14.0  42
## 36    4.8  41
## 37    4.2  39
## 38    5.0  39
## 39    7.0  39
## 40    7.5  39
## 41   13.8  38
## 42    4.1  37
## 43    8.5  37
## 44    5.2  36
## 45   11.9  35
## 46    5.4  33
## 47    8.4  33
## 48    9.3  33
## 49    3.7  32
## 50    9.0  32
## 51   11.2  32
## 52    6.0  31
## 53    7.1  31
## 54    7.3  30
## 55    7.6  30
## 56    5.7  29
## 57   10.3  29
## 58   12.8  29
## 59   10.4  28
```

```
## 60   11.1  28
## 61    9.7  26
## 62   11.6  26
## 63   13.4  26
## 64   12.6  25
## 65    4.5  24
## 66   10.1  23
## 67   10.2  23
## 68    9.8  22
## 69    9.9  22
## 70    3.9  21
## 71   10.0  21
## 72   10.9  21
## 73    9.4  20
## 74   11.5  20
## 75    3.5  18
## 76   12.1  18
## 77    5.8  16
## 78    3.0  15
## 79    7.9  15
## 80   24.9  15
## 81    2.5  14
## 82   12.3  14
## 83    3.6  12
## 84   13.5  12
## 85    8.3  11
## 86   12.0  11
## 87   17.7  11
## 88    9.6  10
## 89    3.1   9
## 90   12.4   9
## 91    2.8   8
## 92    8.8   8
## 93   11.3   8
## 94   14.2   8
## 95   14.9   8
## 96   15.7   8
## 97   10.5   7
## 98   13.0   7
## 99   15.9   7
## 100   3.3   6
## 101  11.0   6
## 102  16.0   6
## 103  16.3   6
## 104   1.4   5
## 105   8.6   5
## 106  11.4   5
## 107  22.3   5
## 108   3.4   4
## 109   7.4   4
## 110  12.5   4
## 111  13.1   4
## 112  13.7   4
## 113   3.2   3
```

```
## 114  14.8   3
## 115  12.9   2
## 116  14.1   2
```

```
hist(CollegeDistance$unemp,
     breaks = 10,
     xlab = "Unemployment",
     ylab = "Frequency",
     main = "Histogram of Unemployment"
)
```



Measurement: This variable represents the unemployment rate in the home state of each student. Level of measurement: Interval Ratio level. Summary: To summarize the unemployment rates, we can use summary statistics. The mean of unemployment in the data set 7.60. The median of unemployment 7.1. The standard deviation of unemployment is 2.76. The range of the data set is 24.9-1.4=23.5. The IQR of the data set is 3. I created a Histogram to get a better visualization of the unemployment variable.

Variable: Distance

```
summary_distance <- CollegeDistance%>%
  summarize(meandistance = mean(distance, na.rm = TRUE),
            mediandistance = median(distance, na.rm = TRUE),
            sddistance = sd(distance, na.rm = TRUE),
            rangedistnace = range(distance, na.rm = TRUE),
            iqrdistance = IQR(distance, na.rm = TRUE))
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
```

```
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
summary_distance
```

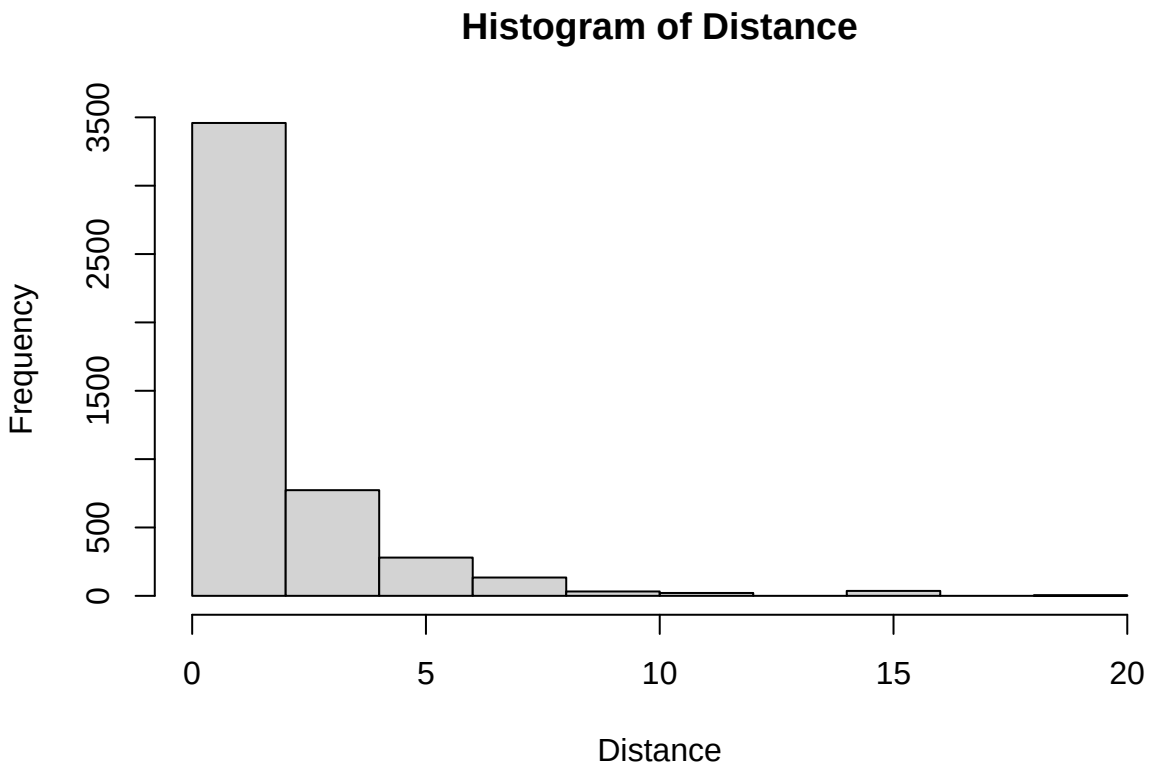
```
##   meandistance mediandistance sddistance rangedistnace iqrdistance
## 1      1.80287              1   2.297128             0         2.1
## 2      1.80287              1   2.297128            20         2.1
```

```
mode_distance <- CollegeDistance%>%
  count(distance)%>%
  arrange(desc(n))
mode_distance
```

```
##    distance   n
## 1       0.5 445
## 2       0.1 420
## 3       1.0 367
## 4       1.5 364
## 5       0.3 308
## 6       0.2 269
## 7       2.0 218
## 8       0.8 202
## 9       0.4 186
## 10      0.7 168
## 11      2.5 163
## 12      4.0 160
## 13      1.2 131
## 14      3.5 127
## 15      3.0 126
## 16      0.6 124
## 17      5.0 121
## 18      0.0  94
## 19      4.5  81
## 20      6.5  47
## 21      1.8  42
## 22      7.0  42
## 23      6.0  41
## 24      2.2  39
## 25      2.6  29
## 26      9.0  26
## 27      1.4  25
## 28      3.8  24
## 29      3.6  20
## 30      3.2  19
## 31      8.0  19
## 32      1.6  18
## 33      1.1  17
## 34      1.7  17
## 35     16.0  17
## 36      2.1  16
## 37      0.9  15
## 38      1.9  15
## 39      2.7  15
```

```
## 40     15.0  15
## 41      1.3  14
## 42      2.8  12
## 43      3.7  12
## 44     12.0  10
## 45      5.6   9
## 46      4.3   8
## 47      5.5   8
## 48      7.8   8
## 49      3.3   7
## 50      4.6   7
## 51      7.6   7
## 52     11.0   7
## 53      6.8   6
## 54     10.0   6
## 55      5.2   5
## 56      2.3   4
## 57     10.5   4
## 58     14.2   4
## 59     20.0   4
## 60      7.5   3
## 61      6.2   2
```

```r
hist(CollegeDistance$distance,
     breaks = 10,
     xlab = "Distance",
     ylab = "Frequency",
     main = "Histogram of Distance"
)
```



Measurement: This variable represents the distance (in miles) between the college and the student's home.

Level of measurement: Interval ratio level. Summary: To summarize the distances, we can use summary statistics. The median of the data set is 1, the mean of the data set is 1.80, the standard deviation of the set is 2.30. the range distance is 20. The IQR distance is 2.1. I created a histogram to get a better visual of the distance variable.

Variable : Income

```
mode_income <- CollegeDistance%>%
  count(income)%>%
  arrange(desc(n))
mode_income
```

```
##   income    n
## 1    low 3374
## 2   high 1365
```

```
freqtable_income <-table(CollegeDistance$income)
freqtable_income
```

```
##
##  low high
## 3374 1365
```

Measurement: This variable represents is the family income above USD 25,000 per year? Level of measurement: Nominal Variable. Summary: To summarize the family income above USD 25,000 per year, we can use a frequency table to count the number of students from each state. Since this is a Nominal variable, we can only calculate the mode.

Variable: Wage

```
summary_wage <- CollegeDistance%>%
  summarise(meanwage = mean(wage, na.rm = TRUE),
            medianwage = median(wage, na.rm = TRUE),
            sdwage = sd(wage, na.rm = TRUE),
            rangewage = range(wage, na.rm = TRUE),
            iqrwage = IQR(wage, na.rm = TRUE))
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
summary_wage
```

```
##   meanwage medianwage   sdwage rangewage  iqrwage
## 1 9.500506       9.68 1.343067      6.59 1.299999
## 2 9.500506       9.68 1.343067     12.96 1.299999
```

```
mode_wage <- CollegeDistance%>%
  count(wage)%>%
  arrange(desc(n))
mode_wage
```

```
##     wage   n
```

```
## 1   8.89 620
## 2   9.92 436
## 3   9.64 335
## 4  10.28 270
## 5   9.96 260
## 6   7.54 256
## 7  11.62 223
## 8  10.15 188
## 9  12.15 176
## 10 10.04 153
## 11  7.33 120
## 12  9.68 118
## 13  8.26 105
## 14 10.51 103
## 15 10.81 103
## 16  8.09 100
## 17  7.18  92
## 18  7.35  89
## 19 11.08  89
## 20  9.76  72
## 21  8.85  69
## 22 10.03  66
## 23 12.96  65
## 24  8.32  62
## 25  7.49  59
## 26  9.90  56
## 27  8.41  55
## 28  7.40  51
## 29 11.56  42
## 30  9.07  39
## 31  7.69  36
## 32  8.10  36
## 33  9.29  36
## 34  9.55  36
## 35  9.73  34
## 36  6.59  31
## 37  7.04  21
## 38 11.37  16
## 39  8.65  13
## 40  7.09   7
## 41  9.50   1
```

```
hist(CollegeDistance$wage,
     breaks = 10,
     xlab = "Wage",
     ylab = "Frequency",
     main = "Histogram of Wage"
)
```

Measurement: This variable represents the hourly wage of each student's part-time job (if applicable). Level of measurement: Interval Ratio level. Summary: To summarize the wages, we can use summary statistics. The mean wage of this data set is 9.50. The median of the data set 9.68. The standard deviation of wage data set is 1.34. The Range of the data set is 12.96-6.59=6.37. The IQR is 1.29. I created a histogram to show the frequency of wage from this data set.

Questions

What is the relationship between level of income (measured as high/low) and whether or not a respondent owns their own home? First, explain what you think this relationship might be. Then, build a visualization to explore that hypothesis. Then estimate the relationship numerically (both in terms of pattern and strength) and comment on its existence in the broader population.

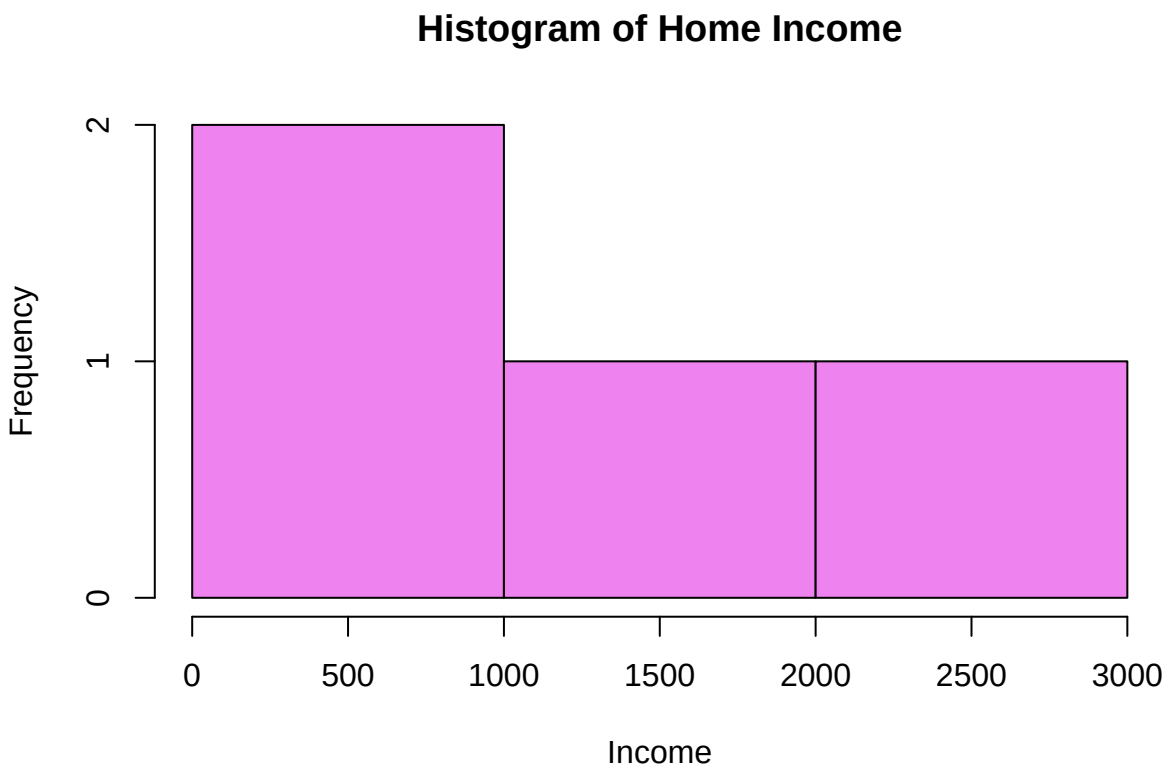
```
freqtable_home_income <- table(CollegeDistance$home, CollegeDistance$income)
freqtable_home_income
```

```
##
##        low high
##   no   721  131
##   yes 2653 1234
```

```
col_proportions <- prop.table(freqtable_home_income, 2)
print(col_proportions)
```

```
##
##             low      high
##   no  0.2136929 0.0959707
##   yes 0.7863071 0.9040293
```

```
hist(freqtable_home_income, main = "Histogram of Home Income", xlab = "Income", ylab = "Frequency", col
```



Chi Square Test

Step 1: Assumptions

1.This data is based off a random sample from college graduates who graduated from high school in 1986. (Does not state anywhere but Professor Freeman told me)

2. The two variables, income and home are independent.

3. Both variables, home and inome are measured nominally.

Step 2: Hypotheses

H0: There is no association between if there family income is above 25,000 USD and if there family owns a home. The two variables are independent. H0 : fe = fo, x2= 0 H1:There is an association between if there family income is above 25,000 USD and if there family owns a home. The two variables are dependent. H1: fe fo, x2> 0

Step 3: Sampling Distribution and Critical Region df=(r-1)(c-1) = (2-1)(2-1)=1, so degrees of freedom = 1 alpha = 0.05 Use appendix C to find critical value X(critical) =3.841

Step 4: Test statistic

```
chi_square <- chisq.test(freqtable_home_income)
print(chi_square)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  freqtable_home_income
## X-squared = 90.536, df = 1, p-value < 2.2e-16
```

Step 5: Make a decision and interpret the decision

Because our test-statistic of 90.536 does fall in the critical region defined by alpha = .05. We reject the null hypothesis that there is no association between if there family income is above 25,000 USD and if there family owns a home. The two variables are independent.

The probability of observing a pattern of differences like the one reflected in this sample if the null hypothesis was true is less than .05, therefore we can conclude that there is a statistically significant relationship between if there family income is above 25,000 USD and if there family owns a home in the broader population.

This test was conducted with a random sample from college graduates who graduated from high school in 1986.

```
strength <- cramerV(freqtable_home_income)
strength
```

```
## Cramer V
##   0.1388
```

Cramer's V is a measure of association between two categorical variables in a contingency table. It ranges between 0 and 1, where 0 indicates no association and 1 indicates a perfect association. Since Crakers V value of .1388 suggests a weak association between if there family income is above 25,000 USD and if there family owns a home in the broader population. This means that there is some degree of association between these two variables, but the strength of the association is relatively weak. It indicates that the ownership of a home is not strongly determined by whether the family income is above $25,000 USD or not.

Estimate the true population mean distance a respondent lives from a 4-year college.

Confidence Interval

```
distancemean <- mean(CollegeDistance$distance)
distancemean
```

```
## [1] 1.80287
```

```
sddistance <- sd(CollegeDistance$distance)
sddistance
```

```
## [1] 2.297128
```

```
sample_size_distance <- length(CollegeDistance$distance)
sample_size_distance
```

```
## [1] 4739
```

```
confidence_level <- 0.95
confidence_level
```

```
## [1] 0.95
```

```
SE <-sddistance/sqrt(sample_size_distance)
SE
```

```
## [1] 0.03336889
```

```
MOE <- qnorm(1 - (1 - confidence_level) / 2) * SE
MOE
```

```
## [1] 0.06540183
```

```
lower_bound <- distancemean - MOE
lower_bound
```

```
## [1] 1.737468
```

```
upper_bound <- distancemean + MOE
upper_bound
```

```
## [1] 1.868272
```

Based on the results conducted from the Confidence Interval Test (confidence level 95%): The sample mean distance a respondent lives from a 4-year college is estimated to be approximately 1.80287 units. The sample standard deviation of the distances is approximately 2.297128 units. The sample size for the distance variable is 4739. The standard error (SE) is calculated to be approximately 0.03336889. The margin of error (MOE) is approximately 0.06540183, obtained by multiplying the SE with the critical value (qnorm) corresponding to the desired confidence level.The lower bound of the confidence interval is approximately 1.737468, obtained by subtracting the MOE from the sample mean. The upper bound of the confidence interval is approximately 1.868272, obtained by adding the MOE to the sample mean.

With 95% confidence, we can estimate that the true population mean distance a respondent lives from a 4-year college falls within the range of approximately 1.737468 to 1.868272 units. This means that we are 95% confident that the actual population mean distance lies within this interval,1.737468 to 1.868272 units, based on the provided sample data.

What is the average difference in county unemployment rate for those respondents who earn above 25,000 dollars a year compared to those respondents who earn less than 25,000 dollars a year? Is this difference likely to be significant in the broader population?

Two Sample T test

Step 1: Assumptions

1. This data is based off a random sample from college graduates who graduated from high school in 1986. (Does not state anywhere but Professor Freeman told me)
2. CLT Applies, N1 + N2 > 100
3. Level of measurement for unemployment is Interval Ratio and for income is Nominal.
4. The samples are Indepdendent.

Step 2: Hypotheses H0: There is no significant difference in the county unemployment rate between respondents earning above 25,000 USD and those earning less than 25,000 USD in the broader population. H0: u1 = u2 H1: There is a significant difference in the county unemployment rate between respondents earning above 25,000 USD and those earning less than 25,000 USD in the broader population. H1: u1 != u2

Step 3: Sampling Distribution and Critical Region T distribution two tailed test a = .05 + or - 1.96 df = 2582.4

Step 4: Test Statistic

```
unemp_income <- t.test(CollegeDistance$unemp~CollegeDistance$income, alternative = "greater", conf.level
unemp_income
```

```
##
##  Welch Two Sample t-test
##
## data:  CollegeDistance$unemp by CollegeDistance$income
## t = 5.4513, df = 2582.4, p-value = 2.737e-08
## alternative hypothesis: true difference in means between group low and group high is greater than 0
## 95 percent confidence interval:
##  0.3328298       Inf
## sample estimates:
##  mean in group low mean in group high
##           7.734529           7.257802
```

```
unemp_income$statistic
```

```
##        t
## 5.451317
```

```
unemp_income$estimate
```

```
##  mean in group low mean in group high
##           7.734529           7.257802
```

```
unemp_income$p.value
```

```
## [1] 2.736674e-08
```

Step 5: Make a decision and interpret the decision Because our test statistic of 5.45 does fall into the critical region defined by alpha=0.05, we reject the null hypothesis that there is no significant difference in the county unemployment rate between respondents earning above 25,000 USD and those earning less than 25,000 USD in the broader population. The probability that we would observe a sample mean of group low is 7.73 years and group high is 7.26 years if the null hypothesis was true is less than 1. Therefore, we conclude that there is a significant difference in the county unemployment rate between respondents earning above 25,000 USD and those earning less than 25,000 USD. This test was conducted with a random sample from college graduates who graduated from high school in 1986.

What is the population proportion of women who own their own home? What do you hypothesize about this proportion given the answer to the earlier question about home ownership and income?

Confidence Interval

```
female_data <- CollegeDistance %>%
  filter(gender == "female")
```

```
alpha <- .05
own_home_count <- sum(female_data$home == "yes")
own_home_count
```

```
## [1] 2098
```

```
total_female_count <- length(female_data$home)
total_female_count
```

```
## [1] 2600
```

```
sample_proportion <- own_home_count / total_female_count
sample_proportion
```

```
## [1] 0.8069231
```

```
standard_error <- sqrt((sample_proportion * (1 - sample_proportion)) / total_female_count)
standard_error
```

```
## [1] 0.007740956
```

```
confidence_level <- 0.95
margin_of_error <- qnorm(1 - (1 - confidence_level) / 2) * standard_error
margin_of_error
```

```
## [1] 0.01517199
```

```
lower_boundincome <- sample_proportion - margin_of_error
lower_boundincome
```

```
## [1] 0.7917511
```

```
upper_boundincome <- sample_proportion + margin_of_error
upper_boundincome
```

```
## [1] 0.8220951
```

Based on the results conducted from the Confidence Interval Test (confidence level 95%): The sample proportion of women who own their own home is estimated to be approximately 0.807. The standard error for the proportion is approximately 0.00774. The margin of error is approximately 0.01517, obtained by multiplying the standard error with the critical value (qnorm) corresponding to the desired confidence level.

The lower bound of the confidence interval is approximately 0.792, obtained by subtracting the margin of error from the sample proportion. The upper bound of the confidence interval is approximately 0.822, obtained by adding the margin of error to the sample proportion.

With 95% confidence, we can estimate that the true population proportion of women who own their own home falls within the range of approximately 0.792 to 0.822. This means that we are 95% confident that the actual population proportion lies within this interval,0.792 to 0.822, based on the provided sample data.

Research suggests that the impact of having a father who attended college is likely to increase the years of education that a child earns. The population mean years of education for a student whose father earned a college degree in 1986 was 14.9 years of education. Do the respondents in this dataset whose father earned a college degree differ from the population as a whole?

Chi Square Test

```
freqtable_fcollege <- table(CollegeDistance$fcollege)
dim(freqtable_fcollege)
```

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

```
print(freqtable_fcollege)
```

```
##
##   no  yes
## 3753  986
```

```
col_proportions2 <- prop.table(freqtable_fcollege, 1)
print(col_proportions2)
```

```
##
##  no yes
##   1   1
```

Step 1: Assumption

1.This data is based off a random sample from college graduates who graduated from high school in 1986. (Does not state anywhere but Professor Freeman told me)

2. The two variables Indepdendent
3. Fcollege is measured Nominally.

Step 2: Hypotheses H0: There is no association between education for the respondents in the dataset whose fathers earned a college degree is equal to the population mean of 14.9 years in the broader population. H0: u = 14.9 H1:There is an association between education for the respondents in the dataset whose fathers earned a college degree is different from the population mean of 14.9 years in the broader population. H1: u != 14.9

Step 3: Sampling Distribution and Critical Region df=(r-1)(c-1) = (2-1)(2-1)=1, so degrees of freedom = 1 alpha = 0.05 Use appendix C to find critical value X(critical) =3.841

Step 4 : Test Statistic

```
chi_square2 <- chisq.test(freqtable_fcollege)
print(chi_square2)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  freqtable_fcollege
## X-squared = 1615.6, df = 1, p-value < 2.2e-16
```

Step 5: Make a decision and interpret the decision

Because our test-statistic of 1615.6 does fall in the critical region defined by alpha = .05. We reject the null hypothesis that There is no association between education for the respondents in the dataset whose fathers earned a college degree is equal to the population mean of 14.9 years in the broader population.

The probability of observing a pattern of differences like the one reflected in this sample if the null hypothesis was true is less than .05, therefore we can conclude that there is a statistically significant relationship between education for the respondents in the dataset whose fathers earned a college degree is equal to the population mean of 14.9 years.

This test was conducted with a random sample from college graduates who graduated from high school in 1986.

What is the relationship between the respondent's state manufacturing wage and the distance a respondent lives from a 4-year institution? First, explain what you think this relationship might be. Then, build a visualization to explore that hypothesis. Then estimate the relationship numerically (both in terms of pattern and strength) and comment on its existence in the broader population. Why might this matter in terms of recruitment for the admissions office?

I think the relationship between wage, state hourly wage in manufacturing in 1980, and distance, distance from 4-year college (in 10 miles), will be positive. This is because the presence of better job opportunities and economic development near 4-year colleges.

```
plot(CollegeDistance$wage,CollegeDistance$distance,
     main = "Relationship between Wage and Distance",
     xlab = "Wage", ylab = "Distance",
     col = "orange")
```

The scatterplot shows the dispersion of the data points across the range of values for Wage and Distance. Each data point represents an observation from the dataset.

Since there is no specific pattern or trend visible in the scatterplot, it suggests that there is no strong relationship between Wage and Distance. The data points appear to be scattered randomly across the plot, indicating no apparent association between the two variables.

Correlation Step 1: Assumptions

1. The independent variable and dependent variable measured at the interval-ratio level
2. The relationship between the two variables is linear
3. Both variables are normally distributed in the population OR the sample size is > 100 (CLT)
4. This data is based off a random sample from college graduates who graduated from high school in 1986. (Does not state anywhere but Professor Freeman told me)
5. Homoscedasticity

Step 2: Hypotheses H0: There is no association between thestate hourly wage in manufacturing in 1980 and distance from 4-year college (in 10 miles) in the broader population. H0 : Rho = 0 H1:There no association between thestate hourly wage in manufacturing in 1980 and distance from 4-year college (in 10 miles) in the broader population. H1: Rho != 0

Step 3: Sampling Distribution and Critical Value Alpha = 0.05 df = 4737 We use the t-distribution because we do not have population standard deviation data so we'll be using the sample data to estimate it. Therefore, the critical regions for this hypothesis test start at -1.96 and positive 1.96

Step 4: Test statistic

```
correlation <- cor(CollegeDistance$wage, CollegeDistance$distance, method="pearson", use="complete.obs")
correlation
```

```
## [1] -0.0003904288
```

```r
cortest <- cor.test(CollegeDistance$wage, CollegeDistance$distance, method="pearson", use="complete.obs
cortest
```

```
##
##  Pearson's product-moment correlation
##
## data:  CollegeDistance$wage and CollegeDistance$distance
## t = -0.026872, df = 4737, p-value = 0.9786
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02886256  0.02808233
## sample estimates:
##           cor
## -0.0003904288
```

Step 5: Make a decision and interpret the decision

The calculated correlation coefficient (cor) is approximately -0.0003904. This indicates a very weak negative correlation between the two variables. The t-value is approximately -0.026872, and the degrees of freedom (df) are 4737. The p-value associated with the correlation coefficient is 0.9786. This p-value suggests that there is no significant evidence to reject the null hypothesis, which states that the true correlation between wage and distance is equal to 0. In other words, there is no strong evidence to support a significant correlation between the variables. The 95% confidence interval for the correlation coefficient is approximately -0.02886256 to 0.02808233. Since this interval contains 0, it further supports the lack of a significant correlation between wage and distance.

Based on the analysis, there is no substantial evidence to suggest a meaningful correlation between state hourly wage in manufacturing in 1980 and the distance from a 4-year college. The correlation coefficient is close to zero, indicating a very weak and negligible relationship between the variables.

Relating back to the scatter plot, the lack of a clear linear or nonlinear pattern in the scatterplot aligns with the previous interpretation of the weak correlation coefficient close to zero.

Conclusion

In conclusion, this research report aimed to understand the patterns of social life among college graduates who attended college over 40 years ago, specifically focusing on the 1986 cohort. By analyzing data from the CollegeDistance dataset, I explored various variables such as homeownership, unemployment rate, distance from a 4-year college, income level, and state manufacturing wages and if fathers are college graduates. Our objective was to compare the data from 1986 with the current data, identifying changes in patterns of social life and informing more effective resource allocation for student recruitment at Hobart and William Smith Colleges (HWS).

Throughout this analysis, I conducted a series of tests and analyses to gain insights into the residential patterns of college graduates from the 1986 cohort. Our research questions encompassed multiple test statistics, allowing me to explore the relationships between variables and evaluate their significance in understanding the distribution of college graduates from that era.

Firstly, I examined the association between income and homeownership among the 1986 college graduates, speculating on its potential connection and extending our analysis to the broader population. I discovered that there is a statistically significant relationship between if there family income is above 25,000 USD and if there family owns a home in the broader population. Additionally, I estimated the true population mean distance from a 4-year college for the respondents and analyzed the average difference in county unemployment rates based on income levels, from this I discovered that I am 95% confident that the actual population mean distance lies within this interval,1.737468 to 1.868272 units, based on the provided sample data. From there I went to test if the average difference in county unemployment rate for those respondents who earn above 25,000 dollars a year compared to those respondents who earn less than 25,000 dollars a year, from this I

found that there is a significant difference in the county unemployment rate between respondents earning above 25,000 USD and those earning less than 25,000 USD. Exploring the population proportion of women who own their homes, I also discussed the hypothesized relationship between income, homeownership, and gender, from this test statistic I discovered that I am 95% confident that the actual population proportion lies within this interval,0.792 to 0.822, based on the provided sample. Moreover, I investigated the impact of a father's college attendance on the educational attainment of their kids, assessing potential differences between the respondents in our dataset and the overall population, from that test I discovered that there is a statistically significant relationship between education for the respondents in the dataset whose fathers earned a college degree is equal to the population mean of 14.9 years. Lastly, I explored the relationship between state manufacturing wages and the distance respondents lived from a 4-year institution, considering its implications for recruitment strategies at the admissions office, from this test I found The 95% confidence interval for the correlation coefficient is approximately -0.02886256 to 0.02808233, there is no strong evidence to support a significant correlation between the variables.

I will now break it down so it is easier to understand (real people words):

Regarding question 1, *What is the relationship between level of income (measured as high/low) and whether or not a respondent owns their own home? First, explain what you think this relationship might be. Then, build a visualization to explore that hypothesis. Then estimate the relationship numerically (both in terms of pattern and strength) and comment on its existence in the broader population.*

I examined the relationship between income and homeownership among college graduates from the 1986 cohort and extended the analysis to the broader population. Then found a statistically significant relationship between having a family income above $25,000 USD and owning a home in the broader population.

Regarding Question 2, *Estimate the true population mean distance a respondent lives from a 4-year college.*

I estimated the true population mean distance from a 4-year college for the respondents and analyzed the average difference in county unemployment rates based on income levels. Then concluded, with 95% confidence, that the actual population mean distance falls within the interval,1.737468 to 1.868272 units.

Regarding question 3,*What is the average difference in county unemployment rate for those respondents who earn above 25,000 dollars a year compared to those respondents who earn less than 25,000 dollars a year? Is this difference likely to be significant in the broader population?*

I tested the average difference in county unemployment rates between respondents earning above 25,000 USD and those earning less than 25,000 USD. Then found a significant difference in the county unemployment rate between the two income groups.

Regarding question 4, *What is the population proportion of women who own their own home? What do you hypothesize about this proportion given the answer to the earlier question about home ownership and income?*

I explored the population proportion of women who own their homes and discussed the hypothesized relationship between income, homeownership, and gender. Then calculated a test statistic and found, with 95% confidence, that the actual population proportion falls within the interval, 0.792 to 0.82.

Regarding Question 5, *Research suggests that the impact of having a father who attended college is likely to increase the years of education that a child earns. The population mean years of education for a student whose father earned a college degree in 1986 was 14.9 years of education. Do the respondents in this dataset whose father earned a college degree differ from the population as a whole?*

I investigated the impact of a father's college attendance on the educational attainment of their children. Then compared the respondents in their dataset to the overall population and discovered a statistically significant relationship between education for the respondents whose fathers earned a college degree, which is equal to the population mean of 14.9 years

Regarding Question 6, *What is the relationship between the respondent's state manufacturing wage and the distance a respondent lives from a 4-year institution? First, explain what you think this relationship might be. Then, build a visualization to explore that hypothesis. Then estimate the relationship numerically (both*

in terms of pattern and strength) and comment on its existence in the broader population. Why might this matter in terms of recruitment for the admissions office?

I examined the relationship between state manufacturing wages and the distance respondents lived from a 4-year institution. Then I considered the implications for recruitment strategies at the admissions office. Based on the analysis, I found that there is no strong evidence to support a significant correlation between the variables. The 95% confidence interval for the correlation coefficient is approximately -0.02886256 to 0.02808233

Overall, this research sheds light on the evolving patterns of social life among college graduates over the past four decades. The findings derived from this analysis will provide crucial insights to the admissions office at HWS, facilitating a comparative analysis of current data with that from over 40 years ago. These insights will enable the identification of shifts and changes in patterns of social life since 1986, empowering the admissions office to allocate their resources more effectively for recruitment efforts in areas with a high potential for attracting new students. By understanding the dynamics of social life among college graduates from different eras, HWS can adapt and evolve to meet the needs and aspirations of prospective students, leading to a vibrant and inclusive community of learners. I suggest that the HWS admissions stay up to date with how their graduates do and where they end up. Knowing this information, can lead schools to target their alumni and potentially get more prospect applicants. All of this can lead to further engagement with incoming classes to express if the previous student bodies were successes. I would also suggest that HWS admissions to consider initiaitves when encouraging parental involvement and provide resources for parents who did not have the opportunity to attend college, to support their children's educational aspirations. With this knowledge and recommendations, HWS can update their recruitment strategies and effectively engage with prospective students, ensuring a welcoming and diverse community for years to come.