**MSBA 5303.15884 - Programming for Analytics**

**Analyzing Public Use Microdata Samples (PUMS) To Find Student Enrollment Trends in Oklahoma**

**CODE-BENDERS**

Rohit Chandrakant Deshpande

**December 4, 2019**

Table of Contents

[Executive Summary 3](#_Toc26368593)

[Introduction 4](#_Toc26368594)

[Project Details 5](#_Toc26368595)

[Data Understanding 6](#_Toc26368596)

[Data Preparation 9](#_Toc26368597)

[Data Visualization and Exploration 13](#_Toc26368598)

[Analysis and Results 27](#_Toc26368599)

[Conclusion 37](#_Toc26368600)

[Recommendations 38](#_Toc26368601)

[Limitation of Study 40](#_Toc26368602)

[Further Research 40](#_Toc26368603)

[References 41](#_Toc26368604)

[Appendix 42](#_Toc26368605)

# 

# **Executive Summary**

Higher education institutions are constantly being faced with the daunting challenge to keep up with their long-established business models. Therefore, there is a constant need to improve, and increase the number of students not only to generate revenue, but most importantly sustainability

University of Central Oklahoma (UCO) has encountered periods where the enrollment numbers were good and there were times when the figures dropped. Prior to 2014, record on student enrolment at UCO had remained steady at about 17,000 for 4 consecutive years. However, in 2014, enrollment began to decline, and this has noticeably plummeted to the present figure at approximately 14,000. The enrollment rate at UCO has declined by 12.4% since 2015. The university and the department of the Office of Institutional Research (OIR) are unable to pin-point the source and cause of this problem.

Enrolment into tertiary institutions is considered for a varied number of reasons, as it is widely accepted as an optional stage of formal learning following secondary education. The decision to learn via post-secondary education is based on some factors such as, disposable income, time, age, course preference and many others. Our mission on this project is to analyze Oklahoma State Public Use Micro-Data Samples (PUMS) and the findings of this research will be used by the University’s Office of Institutional Research to understand the institution’s market which comprises of 91.7% of Oklahoma residents.

The first approach was to analyze UCO’s peer institutions within Oklahoma to determine if they are similarly experiencing this decline issue. This will enable us to decipher the appropriate approach needed to solve the issue. The next phase of the project was channeled to analyzing carefully selected components like gender, educational attainment, employment status to mention a few from the Oklahoma residents’ sample data to enable us to understand the factors that influence the decision of individuals to enroll in higher institution.

With this research project we are seeking to explore if there are any changes in the population variables that might likely have an impact on enrollment figures at UCO. Through analyzing our data, we will find the correlation between the dependent variable which is enrollment and the selected independent variables. With our findings, UCO will be able to take actionable steps based on the solutions we tender, its impact on their bottom line and more importantly, estimate the profitability of potential investments in actualizing an increased enrollment. We hope our findings will help formulate ideas that will lead to an increase in student enrollment at UCO.

At the completion of this project, our results prove that unemployment play a major factor in higher institution enrollment.

# **Introduction**

The University of Central Oklahoma (UCO), located in Edmond, is an institution that is ranked one of the top public regional universities in the state of Oklahoma. Founded in 1890, it is the third largest in Oklahoma and exists to help students learn by providing transformative educational experience so that it’s students may become productive, ethical, and engaged citizens; and become leaders serving the global community.

UCO currently has 14,821 students enrolled and about 1,089 faculty–539 full-time and 550 adjunct faculty. It offers 126 programs and possess around 74 buildings with 5 of the buildings in downtown Oklahoma City. The institution contributes to the intellectual, cultural, economic, and social advancement of the communities and individuals it serves.

The Office of Institutional Research (OIR) is a department at UCO. The mission of this office is to provide data, extensive research, perform analysis and proffer actionable insights to be utilized by decision-makers at all levels for appropriate actions to be taken and to aid the university in attaining its objectives.

The OIR had detected an incessant decline in student’s enrollment at UCO in recent years. Our team, Codebenders will analyze sample data of Oklahoma residents to seek for trends in enrollment and understand UCO’s market in order to proffer productive solutions to its investors and stakeholders and submit findings and recommendations to the OIR.

# **Project Details**

The data that analyzed is the Oklahoma Population Records from sets of years ranging from 2008-2012, 2015, 2016 and 2017 respectively, which is being pulled from “The American Community Survey Public Use Micro-Data Sample (PUMS)” files obtained by the United States Census Bureau. Our data consists of 284 variables and a total of 296,700 entries. We would be comparing two-time frames - (2008 to 2011); this period covers the duration when there was a substantial and apparent increase in enrollment. The subsequent range of years to be considered, will be- (2015 to 2018), as it reflects the period when the enrollment began to decline. We utilized diverse tools and software whilst undergoing this project. Python was used to clean, construct and merge data. After data preparation, we had about 8 components which would turn out to be our variables. In understanding the relationship between the variables, we used Python do statistical analysis and logistic regression of the data. For visualization, we used Tableau and Python to display in pictorial and graphical format our findings and the correlation between the variables. The dataset is a .csv file, opened through Excel.

We only have some possible reasons for this problem. One of them is the decrease in employment rate. According to Nate Johnson’s 2015 study, employment rate has a direct relationship with college enrollment rate. This is likely so as, low employment rate signifies limited opportunity to afford college tuition. Furthermore, a more probable reason could be an increase in tuition cost, as this is evident at UCO. With the advancement of technology, and as we are now in the digital age, a major reason for the decline in enrollment may be attributed to the alternative methods of learning. The rise in ‘e-learning’ seems like the go to option for prospective UCO students. Those within the age bracket to attend college seek comfort, convenience and flexibility which many online learning platforms are offering. Examples include, Coursera, Udemy, YouTube and many more.

Our proposed solutions will be highly dependent on our findings. A more pragmatic approach for UCO to tackle the decline in enrollment will involve a total overhaul of the entire system by re-strategizing. There might be a need to go back to the drawing board and re-design the major courses offered to suit their potential student’s needs, increase employment prospects and more importantly ensure that they are competitive. Secondly, revisiting their advertising campaigns to focus on the right target market, will generate more leads and eventually lead to an increase in enrollment. UCO, amongst its peers has very outstanding programs, for instance, its business college is accredited by the AACSB. Our group is suggesting that more light should be shed on the institution’s achievements to attract more students. In conclusion, UCO should put a substantial effort into seeking private funding to be able to offer more scholarships to its students in the future.

# Data Understanding

**Data Source**

The dataset gathered and analyzed is Public Use Microdata Sample (PUMS) data recorded as part of the American Community Survey (ACS) through the United States Census Bureau. The data acquired is the population census data of Oklahoma residents from 2008-2012 and 2013-2017. Just as the name PUMS indicates, the data is public so access to the data is available via the Bureau’s File Transfer Protocol (FTP) site (www.census.gov/programs-surveys/acs/data/pums.html) via [www.census.gov](http://www.census.gov). The FTP site was accessed through American Community Survey (ACS) - PUMS DATA page. By selecting the folders “2012” then “5-Year” respectively we found the zip file, “csv\_pok.zip”, which contains the Population of Oklahoma dataset from 2008-2012 in csv file format. Data for the years 2013-2017 were found using the same method.

**Data Description**

The PUMS data comprises of information from the files:

1. “ss12pok.csv” is the file of Oklahoma Population Data between 2008-2012. This dataset contains 185,322 records and 290 fields.
2. “psam\_p40.csv” is the file of Oklahoma Population Data between 2008-2012. This dataset contains 185,817 records and 286 fields.

The 290 fields could not all be analyzed for this particular project, so we chose 8 specific fields and they include:

* SERIALNO – a 13-character integer that represents the Year of the record (1st four digits) and the Housing unit. In this case, we were only concerned with the year.
* SCHG (School Grade Level Attending) – an integer that represents the current level of educational enrollment of the individual.
* SCHL (School Grade Level Attained) - an integer that represents the highest level of educational attainment of the individual.
* ESR (Employment Status Recode) – an integer that represents the employment status of the individual.
* SCH (School Enrollment) – an integer that represents whether the individual is enrolled in either a public or private school or college or is not enrolled.
* SEX – an integer that represents if the individual is Male or Female
* WAGP (Wages or Salary of last 12 months) – a rounded integer value of the individual’s wages or salary income for the past 12 months.
* AGEP (Age) – an integer that represents the individual’s age.

The Integrated Post-Secondary Education Data System (IPEDS) from Private and Public Institutions we analyzed was a manually produced excel document then was exported to csv format. The data contained 20 records and 4 fields. The fields used were:

* University – representing one of the 5 schools whose data was compared including: University of Central Oklahoma, Northeastern State University, Tulsa Community College, University of Oklahoma and University of Tulsa
* Year – an integer value representing Year of the school record
* Total Enrollment – an integer value that represents the total enrollment recorded for each school
* Average In-State Undergrad Tuition – an integer value the average tuition recorded for each school

The geographic representation data of enrolled students at UCO was obtained from OIR’s factbook for years 2008-2017 via the website: <http://sites.uco.edu/academic-affairs/ir/factbook.asp> The data was originally an excel document before being converted to a .csv format. The file contained 30 records and 3 fields. The fields used were:

* Geographic Representation – denoting the geographic origin of enrolled students either being In-State, Out-of-State or International.
* Year – an integer that represents the of the year of record
* UCO Enrollment Population – an integer representing the enrollment population

**Data Verification**

The fields to be analyzed from the PUMS data were partly completed. Of the 8 fields chosen, 3 of the fields contained Null values. Those fields were SCH (School Enrollment), and WAGP (Wages or Salary of last 12 months). The ACS PUMS Data Dictionary (2017) demonstrates that ‘NaaN’ values in the SCH field represents the individual not being enrolled in any formal form of education because they are simply too young, less than 3 years old. The ESR represents the individual’s employments status, ‘NaaN’ values in this field represent individuals less than 16 years old. The last field containing Null values was WAGP, and in this field ‘NaaN’ values represents individuals less than 15 years old. All other data gotten were error free.

# **Data Preparation**

We realized that the data obtained from varies resources are not readily for us to use, so we need to prepare the data for our further analyze. We used multiple python methods to make our data easy to understand and clear to read

**Renamed Columns**

After importing the data, we found that out there were 290 columns in the first dataset and 286 columns in the second dataset. Meanwhile, most of them have a name that are difficult for people to understand. Thus, the first step we took was to rename the columns. We renamed the column “SERIALNO” to “year”, "SCHG" to "enrollment" , "SCHL" to "Educational attainment", "ESR" to "Employment Status" , "SCH" to "College Type", "SEX" to "gender", "WAGP" to, "wage/salary" and "AGEP" to "age".

**Data Selection**

We understood that most of the columns in our datasets were unnecessary for us. So, we only selected the columns that are valuable for us and those columns are “year”, "enrollment", "Educational attainment", "Employment Status", "College Type", "gender", "wage/salary" and "age". After the data selection, there were only 8 columns left. Most of the analysis work will focus on those columns. We later added our own columns based on those columns.

**Rationale for Inclusion and Exclusion**

The reason why we believed those 8 columns were important to us is because some previous studies indicated that the enrollment rate is highly related to the employment rate, average salary rate and the age. Our early research shows that there is a negative relationship between employment rate and enrollment rate. That is, if the employment rate low, people tend go to colleges to make themselves more compatible with other people. Similar relationship also appeared on average salary and enrollment, if the average salary rate is low, then the college enrollment rate goes up because people aspect to get a better job though study.

**Data Cleaning**

The data we have was not clean data and those columns were still confusing even after we renamed the columns. The reason of this situation is because the dataset used numbers instead of words descriptions to describe the content within the columns. It is impossible for people to understand without the dictionary attached to the dataset. So, we cleaned the data and made it easy to understand. In the “year” column, we changed the serial number like “2008000000022” to “2008” by divide original number by 1000000000 and then change the data type from “float” to “integer”. In enrollment column, we use a function to substitute the numbers in the column. As a result, the “15” “16” in that column were changed to “undergraduate” and “graduate”, the “0” in that column were changed to “not enrolled” and the numbers other than those three were changed to “enrolled but not in college”. We apply similar functions to other columns and made those columns are much more clear, meaningful and easy to understand.

**Data Construction**

We also needed to create our own columns to make the data easy to analyze. For example, we created a variable column called “em” by convert enrollment column content “undergraduate”, “graduate”, “not enrolled”, “enrolled but not in college” to “1”, “2”, “3” and “4”. The reason why we made this conversion is because we later applied some statistic tests based on the enrollment status. The new variables made it easier for us to analyze it. We created addition 2 more variable columns named “ae” and “es” to base on columns “education attainment” and “employment status”. Meanwhile, we believe the college type played a very important role in us analyze, so we also created college type columns to illustrate what kind of college people enrolled. The college type column contents object values such as “public college”, “private college” and “not attended college in 3 months”.

**Derived Attributes**

There are a lot of missing values in this dataset especially in the “salary/wages” column, the way we deal with those missing values is we fill those missing values with “0”. instead of using other more common ways such as “mean” and “median” to fill the missing value. The reason why we fill the missing values with “0” is because the people have a missing value in “wage/salary” column are always “not in the labor force”. Based on this truth, we believe it is much more reasonable to fill the missing value with “0” than use “mean” and “median” wages and salary value. Another column that contains missing values is the “enrollment” we also fill the missing value in that column with “0”.

**Data Integration**

We ended up with 13 columns in our dataset, and those columns are “year”, “enrollment”, “educational attainment”, “employment status”, “college type”, “gender”, “wage/salary”, “age”, “em”, “ae”, “es”, “ct” and “sex”. we believe those columns will help us to analyze the data properly. This dataset we later working on are now contains 13 columns and 371139 rows.

**Data Merging**

We originally had two datasets with the same columns, the only difference between the two datasets is the time span. The first dataset is from 2008 to 2012 and the second dataset is from 2013 to 2017. We merge those two datasets before we did the data cleaning by using the “pd.concat” function.

**Data Formatting**

Utilized pivot table after cleaning each column to verify all information is contained and there was no missing information.

**Dataset Description**

1. The final dataset contains 13 columns and 371139 rows
2. The final dataset contains 5 object columns and they are “enrollment”, “educational attainment”, “employment status”, “CT” and “sex”
3. It also contains 8 integer columns and they are “year”, “college type”, “gender”, “’wage/salary”, “age”, “ae”, “em”, and “es”
4. Three columns “ae”, “em”, and “es” contain dummy variables that we created based on “enrollment”, “Educational attainment” and “employment status”. We created them for apply statistic tests.
5. We find the missing values in “wages/salary” column and fill the missing values with “0”
6. We also find the missing values in “enrollment” column and also fill the missing values with “0”
7. The original datasets only contain integer information that difficult for people to understand, we transfer most of those integer contents into words descriptions and made the data easy to understand.
8. The “enrollment” column contains “undergraduate”, “graduate”,” not enrolled” and “enrolled but not in college”
9. The “educational attainment” columns contains “high sch/GED”, “college degree”, “no formal education” and “no diploma”
10. The “employment status” columns contains “less than 16 years old”, “unemployed”, “not in the labor force” and “employed”
11. The “college type” column contains “private college”, “public college”, “less than 3-year-old” and “not attained college in 3 months”
12. The sex column contains “male” and “female”
13. Pivot table are used after the data cleaning to verify the column contains the information we need.

|  |  |
| --- | --- |
|  |  |

# **Data Visualization and Exploration**

**Comparison Between Universities in Oklahoma**

Our initial approach was to determine if this decline in enrollment was peculiar to only UCO. Our discovery will ascertain if the cause of problem is from the institution or from the Oklahoma population. UCO’s peer group in the state of Oklahoma is divided into four major categories – Regional University System of Oklahoma (RUSO), Oklahoma Research, Oklahoma Community Colleges and Oklahoma Privates. We selected the university with the highest total enrollment in each category:

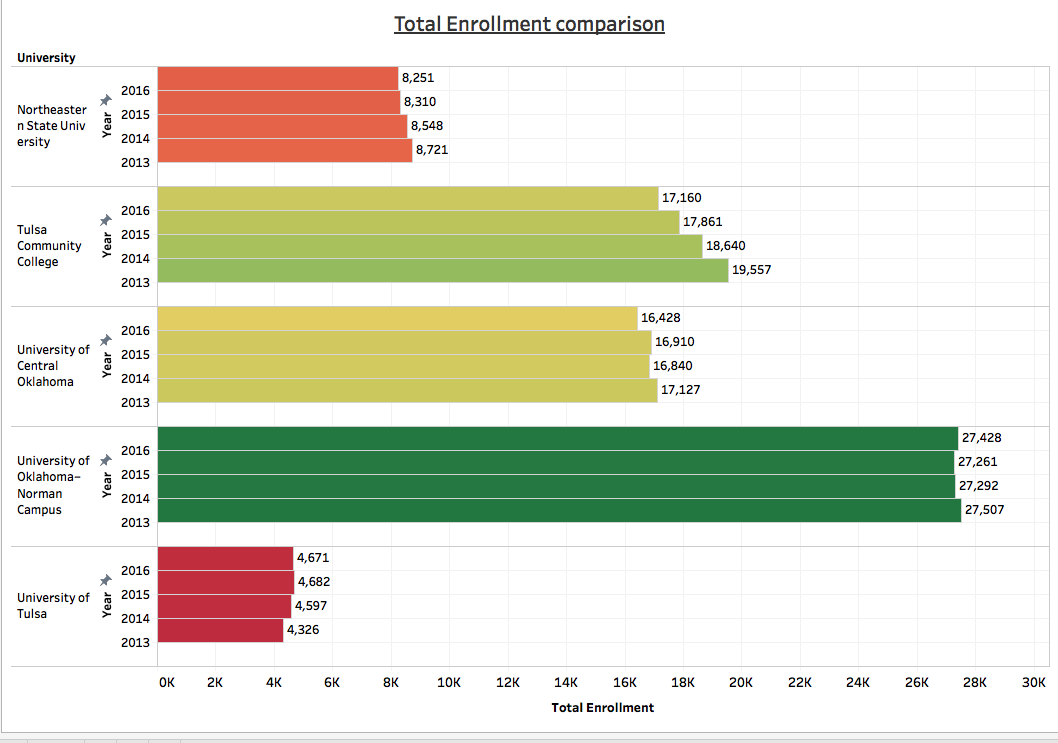
RUSO – Northeastern State University

Oklahoma Research - University of Oklahoma‐Norman Campus

Oklahoma Community Colleges - Tulsa Community College

Oklahoma Privates – University of Tulsa

**Figure 1: Total Enrollment Comparison**



The graph in Figure 1 above proves that there is an evident enrollment issue. Each of the selected universities have experienced enrollment decline. However, enrollment decline is more significant for public institutions. We compared some private universities in Oklahoma to delve further in detecting significant enrollment issues. We selected three universities with the highest total enrollment in the Oklahoma private universities UCO peer group.

**Figure 2: Comparison between Oklahoma Private Universities**

The graph above shows that University of Tulsa and Oral Roberts University are experiencing slight enrollment increase while Oklahoma City University is experiencing enrollment decrease.

The graph in Figure 3 above shows tuition fees increase every year for each institution.

**Figure 3: Comparison between Oklahoma Universities Tuition Fees**

A screenshot of a cell phone

Description automatically generated

**Study of Oklahoma Population**

The dependent variable selected is School Enrollment and the independent variables are Educational Attainment, College Type, Gender, Wage/Salary (total for 12 months), Employment Status and age. Firstly, we examined the variables by School Enrollment and Gender.

**Figure 4: Gender Distribution in Oklahoma**

The number of men and women in Oklahoma population is quite equivalent, with the male representing 49% of the population and women representing 51%.

**School Enrollment Analysis**

Values - Enrolled but not in college (person is in grade school, high school or middle school), Not enrolled (person isn’t enrolled in any form of formal education), Graduate and Undergraduate.

**Figure 5: School Enrollment Distribution**

Figure 5 above shows that most of Oklahoma’s population – 75% are not enrolled in any form of formal education, about 19% are in either grade school, middle high school, about 5% are undergraduates and only about 1% of the population are enrolled in graduate school.

**Figure 6: Higher Education Trend in Oklahoma**

Higher education enrollment began to fall in 2012 and it is yet to be as high as it was in 2011 as shown in Figure 6 above.

The following graphs show a breakdown of enrollment trends by education level – undergraduate and graduate.

A screenshot of a social media post

Description automatically generated**Figure 7: Undergraduate Enrollment Trend in Oklahoma**

Higher education enrollment for undergraduates began to fall in 2012 and it is yet to be as high as it was in 2011 as shown in Figure 7 above.

**Figure 8: Graduate Enrollment Trend in Oklahoma**

A screenshot of a social media post

Description automatically generated

Higher education enrollment for graduates began to fall in 2012 and it is yet to be as high as it was in 2011 as shown in Figure 8 above.

**School Enrollment Analysis by Enrollment Type and Gender**

A close up of a map

Description automatically generated**Figure 9: School Enrollment by gender**

Figure 9 above shows there is a wide gap in higher education enrollment between males and females.

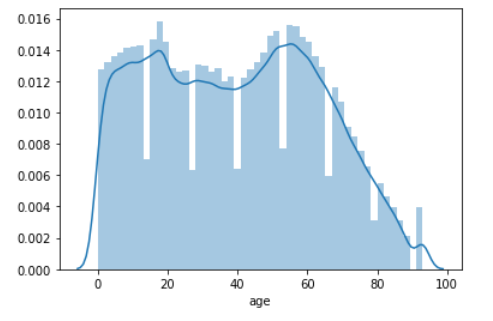
Table 1 displays a pivot table showing the breakdown for school enrollment in Oklahoma by gender over the selected years.

**Table 1**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Enrolled but not in College** | | **Undergraduate** | | **Graduate** | | **Not enrolled** | | **Total population** |
| **Year** | **Male** | **Female** | **Male** | **Female** | **Male** | **Female** | **Male** | **Female** |  |
| **2008** | 348,400 | 334,900 | 67,000 | 98,100 | 13,900 | 15,800 | 1,326,700 | 1,430,500 | 3,635,300 |
| **2009** | 357,100 | 330,000 | 76,300 | 96,600 | 13,000 | 18,800 | 1,376,500 | 1,434,300 | 3,702,600 |
| **2010** | 354,200 | 339,800 | 73,800 | 105,700 | 12,100 | 18,300 | 1,372,400 | 1,443,400 | 3,719,700 |
| **2011** | 362,000 | 327,500 | 86,400 | 111,900 | 13,700 | 18,600 | 1,409,200 | 1,435,700 | 3,765,000 |
| **2012** | 359,800 | 340,600 | 82,600 | 105,900 | 13,200 | 16,900 | 1,371,700 | 1,418,900 | 3,709,600 |
| **2013** | 358,100 | 340,700 | 80,500 | 104,400 | 11,700 | 18,000 | 1,386,300 | 1,414,900 | 3,714,600 |
| **2014** | 360,400 | 342,700 | 89,400 | 99,300 | 11,800 | 17,400 | 1,386,400 | 1,421,900 | 3,729,300 |
| **2015** | 358,100 | 351,900 | 85,300 | 103,200 | 12,500 | 19,100 | 1,382,700 | 1,412,300 | 3,725,100 |
| **2016** | 362,600 | 338,800 | 78,000 | 100,500 | 14,100 | 16,600 | 1,393,200 | 1,398,400 | 3,702,200 |
| **2017** | 364,100 | 341,500 | 75,800 | 102,200 | 11,500 | 16,600 | 1,389,900 | 1,408,900 | 3,710,500 |

**Analysis Of Age Distribution Of The Population**

**Figure 10: Age distribution of Oklahoma population**

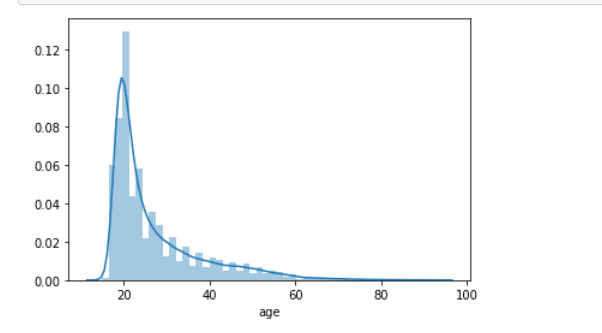
** Table 2**

|  |  |
| --- | --- |
| **Description** | |
| **Count** | 371,139 |
| **Mean** | 39 |
| **Std** | 23 |
| **Min** | 0 |
| **25%** | 19 |
| **50%** | 40 |
| **75%** | 59 |
| **Max** | 93 |

Most of the population’s age range are between 0 – 20 and 40 – 60. The mean age is 39years.

**Age distribution based on School Enrollment categories**

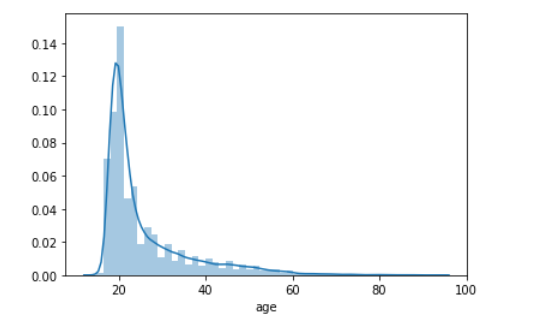
**Figure 11: Age distribution of College students in Oklahoma**

**Table 3**

|  |  |
| --- | --- |
| **Description** | |
| **Count** | 21,265 |
| **Mean** | 27 |
| **Std** | 10 |
| **Min** | 15 |
| **25%** | 20 |
| **50%** | 23 |
| **75%** | 31 |

From Figure 11 above, most college students in Oklahoma have an age range of 15 and 40years, with the mean age being 27 years.

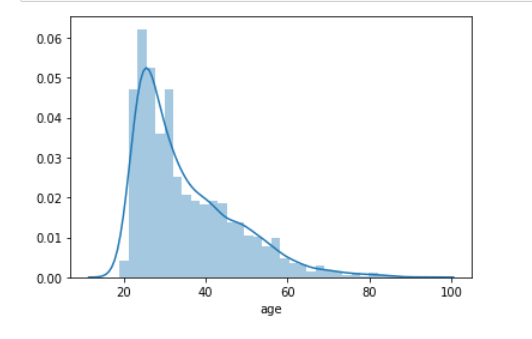
**Figure 12: Age distribution of undergraduate students in Oklahoma**

**Table 4**

|  |  |
| --- | --- |
| **Description** | |
| **Count** | 18,229 |
| **Mean** | 25 |
| **Std** | 10 |
| **Min** | 15 |
| **25%** | 19 |
| **50%** | 22 |
| **75%** | 28 |

From Figure 12 above, most college students in Oklahoma have an age range of 15 and 40years, with the mean age being 25 years.

**Figure 13: Age distribution of graduate students in Oklahoma**

****Table 5**

|  |  |
| --- | --- |
| **Description** | |
| **Count** | 3,036 |
| **Mean** | 35 |
| **Std** | 12 |
| **Min** | 19 |
| **25%** | 26 |
| **50%** | 31 |
| **75%** | 42 |

From Figure 13 above, most college students in Oklahoma have an age range of 19 and 60years, with the mean age being 35 years.

Table 6 below displays a pivot table showing the breakdown of mean age by school enrollment in Oklahoma over the selected years.

**Table 6: Mean Age of Oklahoma population by School Enrollment**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **School Enrollment** | | | |
| **Year** | **Enrolled by not in college** | **Graduate** | **Not enrolled** | **Undergraduate** |
| 2008 | 11 | 34 | 48 | 26 |
| 2009 | 11 | 35 | 47 | 26 |
| 2010 | 11 | 35 | 47 | 26 |
| 2011 | 11 | 35 | 47 | 25 |
| 2012 | 11 | 36 | 47 | 26 |
| 2013 | 11 | 35 | 46 | 25 |
| 2014 | 11 | 34 | 47 | 25 |
| 2015 | 11 | 34 | 47 | 25 |
| 2016 | 11 | 34 | 47 | 25 |
| 2017 | 11 | 34 | 47 | 25 |

**Educational Attainment of the Population**

Values – *no formal education*, *no diploma* (still in grade, middle or high school), *high sch. /GED* (has high school diploma or General Educational Development – GED certificate) and *college degree* (person has at least an associate’s or bachelor’s degree).

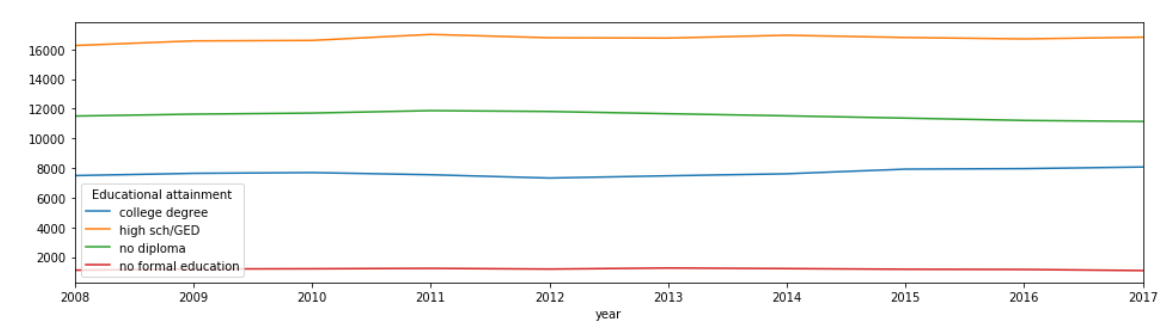
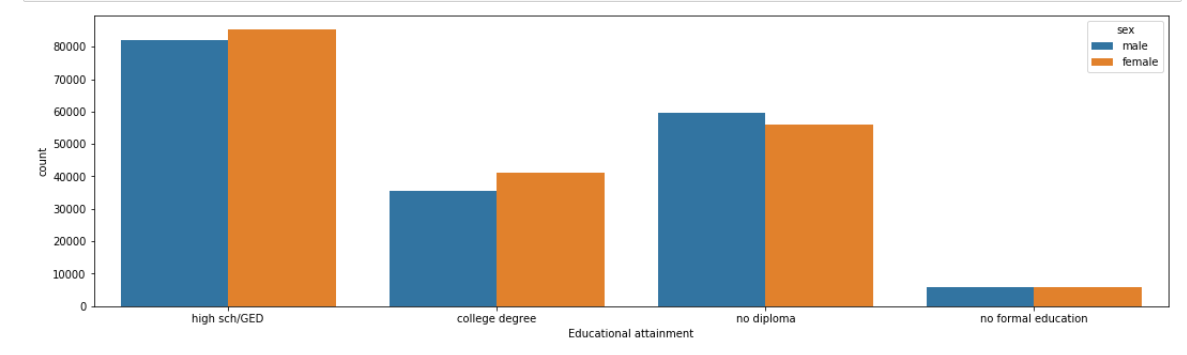
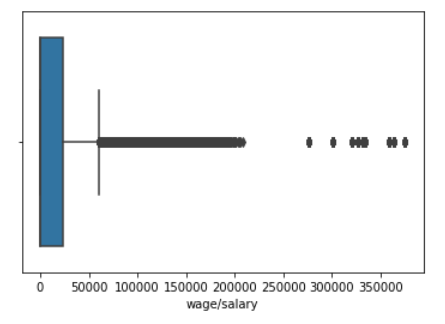
**Figure 14: Educational Attainment of Oklahoma Population**

Figure 14 shows the order of Educational attainment in Oklahoma from highest to lowest. This gives an insight into how educated the Oklahoma’s population is. Most of Oklahoma’s population own a high school diploma, a good number of the population didn’t graduate from high school – hence, no diploma, some received a college degree and the lowest number is that of those with no formal education.

Figure 15 below shows that there are more females than males with high school diploma/GED and college degree. However, there are more men than women with no diploma.

**Figure 15: Educational Attainment by Gender**

**Analysis of Annual Wages And Salary Of The Population**

**Figure 16: Analysis of annual wages and salary of the population Table 7**

|  |  |
| --- | --- |
| **Description** | |
| **Count** | 371,139 |
| **Mean** | $16,505 |
| **Std** | $33,496 |
| **Min** | 0 |
| **25%** | 0 |
| **50%** | 0 |
| **75%** | $24,000 |
| **Max** | $375,000 |

The mean salary of the population across the selected years is $16,505.

**Figure 17: Annual wages and salary trend**

**A close up of a map

Description automatically generated**The lowest mean of wage and salary in the trend is in 2011 as shown in Figure 17. There began to be an increase in mean wage and salary from 2012 and it has not been low as it was in 2011.

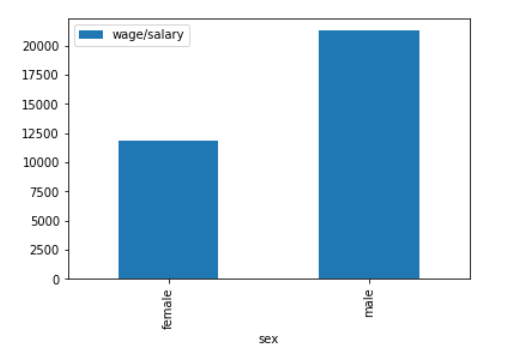
Table 8 below displays a pivot table showing the breakdown of mean, maximum and standard deviation of wages and salary in Oklahoma over the selected years.

**Table 8**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Wages or Salary in 12 months** | | | |
| **Year** | **Mean** | **Max** | **Std** | **Size** |
| 2008 | $16,132 | $334,000 | $31,483 | 36,353 |
| 2009 | $16,066 | $331,000 | $31,506 | 37,026 |
| 2010 | $15,601 | $276,000 | $30,548 | 37,197 |
| 2011 | $15,310 | $301,000 | $31,029 | 37,650 |
| 2012 | $15,612 | $320,000 | $31,799 | 37,096 |
| 2013 | $16,657 | $334,000 | $34,213 | 37,146 |
| 2014 | $16,633 | $327,000 | $33,470 | 37,293 |
| 2015 | $17,594 | $375,000 | $37,178 | 37,251 |
| 2016 | $17,782 | $363,000 | $37,008 | 37,022 |
| 2017 | $17,669 | $359,000 | $35,756 | 37,105 |

**Figure 18: Analysis of wage/salary by Gender**

|  |  |
| --- | --- |
| **Gender** | **Mean Wage** |
| **female** | $11,843 |
| **male** | $21,296 |

****

**Table 9**

**Figure 19: Trends of wage/salary by Gender**

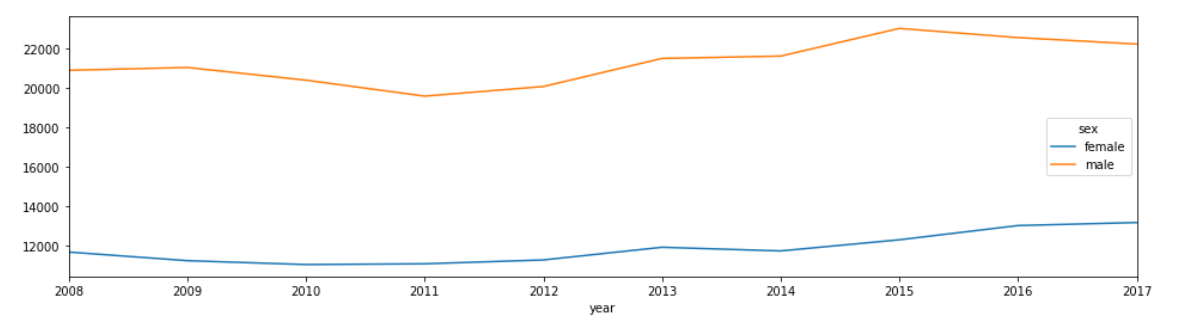
****

Figure 18 and 19 above shows that there is a huge pay gap between men and women in Oklahoma. From the time series analysis, Table 9 shows that the average pay for women is $11,843 and the average pay for men is $21,296. This is a huge difference of $9,453

**Employment Trends of the Population**

Values – *Employed* (civilian or a member of the armed forces that is at work or has a job but not at work), *Not in the labor force* (person is below 16 years, retired or not looking for a job), *Unemployed* (person is looking for a job but hasn’t gotten any). In figure 20 below, about 64% of the population is employed, 34% not in the labor force and 2% unemployed.

A close up of a logo

Description automatically generated**Figure 20: Oklahoma Population by Employment Status**

Table 10 below displays a pivot table showing the breakdown of employment status of the population in Oklahoma over the selected years. Unemployment was at the highest in 2011.

**Table 10**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Employment Status** | | | **Total Population** |
| **Year** | **Employed** | **Not in the labor force** | **Unemployed** |  |
| **2008** | 2,414,700 | 1,148,500 | 72,100 | 3,635,300 |
| **2009** | 2,385,400 | 1,202,200 | 115,000 | 3,702,600 |
| **2010** | 2,382,700 | 1,201,200 | 135,800 | 3,719,700 |
| **2011** | 2,341,500 | 1,296,600 | 126,900 | 3,765,000 |
| **2012** | 2,340,400 | 1,261,400 | 107,800 | 3,709,600 |
| **2013** | 2,362,400 | 1,249,900 | 102,300 | 3,714,600 |
| **2014** | 2,354,200 | 1,281,300 | 93,800 | 3,729,300 |
| **2015** | 2,362,200 | 1,271,700 | 91,200 | 3,725,100 |
| **2016** | 2,323,400 | 1,287,100 | 91,700 | 3,702,200 |
| **2017** | 2,355,800 | 1,264,300 | 90,400 | 3,710,500 |

**Figure 21: Analysis of Employment trends by gender**

**A picture containing screenshot

Description automatically generated**

There are more employed men than women, more women than men are not in the labor force and more men than women are unemployed.

**Comparison between UCO’s Enrollment by Geographic Representation**

We analyzed UCO’s enrollment based on the Geographic origin of the students when they enrolled with the university. Enrollment by geographic representation is separated into 3 categories - In-State**,** Out-of-Stateand International.

**Table 11**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | In-State | Out-of-State | International | Total |
| 2008 | 14245 | 569 | 910 | 15270 |
| 2009 | 14756 | 456 | 880 | 16092 |
| 2010 | 15588 | 482 | 1031 | 17101 |
| 2011 | 15500 | 508 | 1231 | 17239 |
| 2012 | 15337 | 553 | 1321 | 17211 |
| 2013 | 15082 | 601 | 1405 | 17088 |
| 2014 | 14892 | 487 | 1490 | 16869 |
| 2015 | 14993 | 476 | 1449 | 16918 |
| 2016 | 14788 | 458 | 1191 | 16437 |
| 2017 | 14630 | 479 | 1046 | 16155 |

Table 11 above shows enrollment decline is from In-state and International students. In Fall 2018, enrollment for International declined to 838. UCO’s International student enrollment is currently at 756.

# **Analysis and Results**

**T-test** (Significance level: 0.05)

**Title:** Relationship between mean salary of the people in last 12 months versus the enrollment status

**Hypothesis:**

Ho: No Significant difference in the population means of salary and enrollment

Ha: Significant Difference in the population means of salary and enrollment

**Summary statistics:**

A screenshot of a cell phone

Description automatically generated

**Decision:** Reject the Null Hypothesis (Ho)

**Interpretation:** There is a significant difference (p-value =0.000) between the past 12 months’ salary of the people and the total enrollment.

**Chi-square Tests 1** (Significance level: 0.05)

**Title:** Association between Enrollment vs year 2010

**Hypothesis:**

Ho: No association between the year 2010 and the enrollment

Ha: there is an association between the year 2010 and the enrollment

**Summary statistics:**

A receipt on a black background

Description automatically generated

**Decision:** Failed to Reject the Null Hypothesis (Ho)

**Interpretation:** There is no association (p-value =0.458) between the year 2010 and the total enrollment.

**Chi-square Tests 2** (Significance level: 0.05)

**Title:** Association between Enrollment vs year 2011

**Hypothesis:**

Ho: No association between the year 2011 and the enrollment

Ha: there is an association between the year 2011 and the enrollment

**Summary statistics:**

A close up of a receipt

Description automatically generated

**Decision:** Reject the Null Hypothesis (Ho)

**Interpretation:** There is a strong association (p-value =0.001) between the year 2011 and the total enrollment.

**Chi Square Test 3** (Significance level: 0.05)

**Title:** Association between Enrollment vs sex

**Hypothesis:**

Ho: No association between the sex and the enrollment

Ha: there is an association between the sex and the enrollment

**Summary statistics:**

A screenshot of a cell phone

Description automatically generated

**Decision:** Reject the Null Hypothesis (Ho)

**Interpretation:** There is a strong association (p-value =0.000) between the sex and the total enrollment.

**Chi-square Tests 4** (Significance level: 0.05)

**Title:** Association between employment vs school type (private school)

**Hypothesis:**

Ho: No association between the employment in the last 12 months and the admission in the private school

Ha: association between the employment in the last 12 months and the admission in the private school

**Summary statistics:**

A screenshot of a cell phone

Description automatically generated

**Decision:** Reject the Null Hypothesis (Ho)

**Interpretation:** There is a strong association (p-value =0.000) between the employment in the last 12 months and the admission in the private school.

**Collinearity and Logistic Regression**

Collinearity can be seen in between:

* ‘Age’ and ‘school\_1’ (not attended in last 3 months);
* ‘school\_2’ (public school) and ‘school\_1’;
* ‘cat\_employment’ (employment status) and ‘salary’
* ‘cat\_employment’ and ‘educational attainment’

**Heatmap**

A screenshot of a cell phone

Description automatically generated

**Logistic Regression and Results**

After eliminating the collinear variables, we get the below variables in the python code using stats model:

logfit=smf.logit('enrollment ~ salary + age + C(c\_2009)\ +C(c\_2010)+C(c\_2011) + C(c\_2012) + C(c\_2013) \ + C(c\_2014) + C(c\_2015) + C(c\_2016) \ + C(c\_2017)+ C(cat\_Employment) + C(c\_highschool) + \ C(c\_nodiploma) + C(c\_noedu) + C(sex) ',data=c4).fit()

logfit.summary()

**Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | enrollment | **No. Observations:** | 371139 |
| **Model:** | Logit | **Df Residuals:** | 371122 |
| **Method:** | MLE | **Df Model:** | 16 |
| **Date:** | Sun, 01 Dec 2019 | **Pseudo R-squ.:** | 0.1252 |
| **Time:** | 17:46:18 | **Log-Likelihood:** | -71256. |
| **converged:** | False | **LL-Null:** | -81451. |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 0.000 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 0.1245 | 0.037 | 3.348 | 0.001 | 0.052 | 0.197 |
| **C(c\_2009)[T.1]** | 0.0695 | 0.034 | 2.075 | 0.038 | 0.004 | 0.135 |
| **C(c\_2010)[T.1]** | 0.0921 | 0.033 | 2.764 | 0.006 | 0.027 | 0.157 |
| **C(c\_2011)[T.1]** | 0.2023 | 0.033 | 6.187 | 0.000 | 0.138 | 0.266 |
| **C(c\_2012)[T.1]** | 0.1486 | 0.033 | 4.496 | 0.000 | 0.084 | 0.213 |
| **C(c\_2013)[T.1]** | 0.1337 | 0.033 | 4.029 | 0.000 | 0.069 | 0.199 |
| **C(c\_2014)[T.1]** | 0.1457 | 0.033 | 4.405 | 0.000 | 0.081 | 0.211 |
| **C(c\_2015)[T.1]** | 0.1606 | 0.033 | 4.864 | 0.000 | 0.096 | 0.225 |
| **C(c\_2016)[T.1]** | 0.1277 | 03 | 3.825 | 0.000 | 0.062 | 0.193 |
| **C(c\_2017)[T.1]** | 0.0955 | 0.033 | 2.853 | 0.004 | 0.030 | 0.161 |
| **C(cat\_Employment)[T.2]** | -1.6150 | 0.020 | -81.794 | 0.000 | -1.654 | -1.576 |
| **C(c\_highschool)[T.1]** | -1.5138 | 0.026 | -57.683 | 0.000 | -1.565 | -1.462 |
| **C(c\_nodiploma)[T.1]** | -0.7562 | 0.020 | -37.114 | 0.000 | -0.796 | -0.716 |
| **C(c\_noedu)[T.1]** | -22.9788 | 1976.145 | -0.012 | 0.991 | -3896.151 | 3850.193 |
| **C(sex)[T.1]** | 0.2027 | 0.015 | 13.668 | 0.000 | 0.174 | 0.232 |
| **salary** | -2.022e-05 | 4.78e-07 | -42.341 | 0.000 | -2.12e-05 | -1.93e-05 |
| **age** | -0.0408 | 0.000 | -81.630 | 0.000 | -0.042 | -0.040 |

**Best Model for prediction enrollment variable**

Log odds(enrollment=1) = -0.9332 + (0.0591\* year 2009) + (0.0873\* year 2010) + (0.1950\* year 201) + (0.1355\* year 2012) + (0.1069\* year 2013) + (0.1226\* year 2014)+ (0.1318\* year 2015) + (0.0939\* year 2016) + (0.0687\* year 2017) -(1.3954\* employment status in last 12 months) –(0.9932\* schooltype highschool) –(0.0959\* eductional attainment no deploma) +(0.3366\*sex) – (0.0387 \* age)

**Odd Ratio and 95% Confidence Interval**

|  | **OR** | **z-value** | **2.5%** | **97.5%** |
| --- | --- | --- | --- | --- |
| **Intercept** | 0.39 | 0.00 | 0.37 | 0.42 |
| **C(c\_2009)[T.1]** | 1.06 | 0.08 | 0.99 | 1.13 |
| **C(c\_2010)[T.1]** | 1.09 | 0.01 | 1.02 | 1.16 |
| **C(c\_2011)[T.1]** | 1.22 | 0.00 | 1.14 | 1.29 |
| **C(c\_2012)[T.1]** | 1.15 | 0.00 | 1.07 | 1.22 |
| **C(c\_2013)[T.1]** | 1.11 | 0.00 | 1.04 | 1.19 |
| **C(c\_2014)[T.1]** | 1.13 | 0.00 | 1.06 | 1.21 |
| **C(c\_2015)[T.1]** | 1.14 | 0.00 | 1.07 | 1.22 |
| **C(c\_2016)[T.1]** | 1.10 | 0.00 | 1.03 | 1.17 |
| **C(c\_2017)[T.1]** | 1.07 | 0.04 | 1.00 | 1.14 |
| **C(cat\_Employment)[T.2]** | 0.25 | 0.00 | 0.24 | 0.26 |
| **C(c\_highschool)[T.1]** | 0.37 | 0.00 | 0.35 | 0.39 |
| **C(c\_nodiploma)[T.1]** | 0.91 | 0.00 | 0.88 | 0.94 |
| **C(sex)[T.1]** | 1.40 | 0.00 | 1.36 | 1.44 |
| **age** | 0.96 | 0.00 | 0.96 | 0.96 |

**Interpretation:**

* For continuous variable “age”:

*“For each year increase in age, the odds of the person being enrolled changes by a factor of 0.96 or -4%”*

* For categorical variables:

“*For Females, the likelihood of getting enrolled is 1.4 times higher than that for males*”

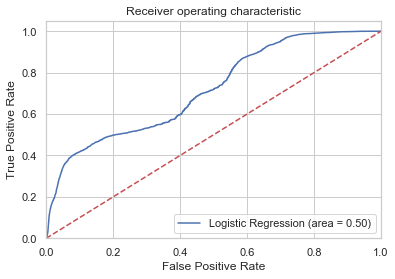
*“Enrollment for employed people in last 12 months is 25% that of the enrollment for unemployed people”*

Similarly, for the other categorical variables such as year, school type and educational attainment we can interpret the odds ratios.

**Accuracy and ROC curve**

Accuracy score: 0.943

ROC curve:



Curve shows that the model is good, though cannot be argued that it’s an excellent model.

**Findings**

1. Public institutions in Oklahoma are majorly affected by enrollment decline issue while among the private universities, some are not being affected.
2. Enrollment for both undergraduate and graduates was increasing but it began to drop in 2012 and it has been decreasing since then.
3. There is significant decline in year 2012 and 2016 for graduate and undergraduate enrollment and also in total population.
4. In grade, middle and high school, there are more males than female enrolled.
5. In college, there are more women than men enrolled.
6. The shift in the mean age shows those who decide to go to college are doing so at an earlier age than before.
7. There are older women than men in Oklahoma.
8. Males either drop out of high school or stop education after getting their high school diploma
9. 2011 is a very significant year.
10. There seems to be a strong correlation between school enrollment, wages/salary and unemployment.
11. There are more educated women than men in Oklahoma
12. There is a huge pay gap between men and women.
13. There are more men than women in the labor force.

# **Conclusion**

According to our research there is a correlation between unemployment rate, wage or salary and higher education enrollment. From our analysis, we can see the enrollment increases when there is a decrease in mean wage/salary and there is an increase in unemployment rate.

When there are enough jobs and there is good pay, Oklahoma residents don’t seem to see the need to attend college but when there are not enough jobs, they tend to go get a degree in order to be able to beat competition to getting a job.

# **Recommendations**

The factors that affect higher education enrollment are economical and are mostly beyond the control of UCO. The university should strive to generate ideas and implement processes to drive the market toward its institution. From our analysis and findings, we tender the following recommendations:

* There is a need to earn income for survival, hence, people need to work. UCO must provide a way for people to combine working and attending school at the same time. The organization should invest in teaching via online platforms whilst making them effective and affordable.
* UCO should organize programs in high schools in order to create awareness on the importance of higher education. The marketing team would need to strategize on how to implement this. The popular way is to talk about how a college degree helps to get a higher pay but since that is not exactly the case anymore, the team will have to think of other ways to emphasize the importance of higher education to an individual and the society. While creating awareness on higher education, UCO must be portrayed has the best place in Oklahoma to get a degree.
* It is a great task to convince people to enroll in college. It is a greater one to ensure they stay in college, that is, to prevents students from dropping out of college. UCO has to listen. One way to do this is to make surveys. Those who have dropped out should be asked for their reason and those currently enrolled should be asked how the university can be improved. There are many reasons students drop out of college such as family issues, too much stress, unprepared for the workload, personal emergency, distractions, lack of advising and so on. UCO must conduct research to know which of the reasons are most common with students that drop out of the university and find ways to provide a solution.
* For international students, many factors also contribute to the enrollment decline in UCO. The major reason is the decline in the number of student visa being granted. This is beyond UCO's control, but UCO can still compete for the little market left. One way is to attract the students while they are still in their home country. UCO along with other Oklahoma universities should partner with educational consultants in countries where their international students come from to organize events to encourage them choose to study in US and in Oklahoma in particular. Other major reason UCO is losing international students is the increase in tuition fees. The dollar is stronger than the currency being used in home countries of most of these students and paying tuition can be very challenging. Federal loans are not available to international students so they attend to pick cheaper schools they can easily afford.
* UCO's used to be a cheap school but the government funding for the school as been reduced and the university had to increase tuition fee in order to be able to continue running the school. This is understandable but a good payment plan has to be set up for both US and international students in order to encourage them to enroll regardless of the tuition fees.

UCO also had to work hard to raise funds through donations from private individuals and

organizations to help run the university. One way could be to put students as the face of

these campaigns, using emotional appeal to draw the attention of wealthy people in the

society.

# **Limitation of Study**

• The 2018 PUMS data was not published until mid-November. It was too late to use it as we have gone far with our project by then.

• Data on UCO's peer institutions is not up to date.

• When gathering information from other universities institutional research site, we discovered a lot of data is missing and most of the universities didn't have up-to-date information.

# **Further Research**

Further study to be carried out:

* Study of the economy on election years
* Study on reason males drop out of high school
* Study on Oklahoma private institutions

# **References**

Ashley S. (2018, June 21): No Bottom Yet in 2-Year College Enrollments. Inside Higher Ed

Retrieved from [https://www.insidehighered.com/news/2018/06/21/community-college-enrollment-rates-expected- keep-falling](https://www.insidehighered.com/news/2018/06/21/community-college-enrollment-rates-expected-%20keep-falling)

Ben F. (2017, August 13) As rural Oklahoma enrollment declines, so does funding The Oklahoman. Retrieved from <https://oklahoman.com/article/5559744/as-rural-oklahoma-enrollment-declines-so-does-funding>

E-blog (2016, November 22): Solutions for Declining Enrollment. PowerSchool. Retrieved from <https://www.powerschool.com/resources/blog/solutions-declining-enrollment/>

Michael M., Michael L., and Kathleen M. (2017August 23): A Lost Decade in Higher Education Funding. Center on Budget and Policy Priorities. Retrieved from <https://www.cbpp.org/research/state-budget-and-tax/a-lost-decade-in-higher-education-funding>

Nate J. (2015, August 2015): The Unemployment-Enrollment Link. Inside Higher Ed. Retrieved from [https://www.insidehighered.com/views/2015/08/27/unemployment-rate-community-college- enrollments-and-tough-choices-essay](https://www.insidehighered.com/views/2015/08/27/unemployment-rate-community-college-%20enrollments-and-tough-choices-essay)

Richard V. (2018, July 5): Why Enrollment Is Shrinking At Many American Colleges. Forbes

Retrieved from <https://www.forbes.com/sites/richardvedder/2018/07/05/academic-deserted-villages/#90d18565121b>

Sarah M. And Stephen S. (2018, October 17): A look at the correlation between enrollment and unemployment. The Horizon. Retrieved from [https://www.iushorizon.com/22822/news/a-look-at-the-correlation-between-enrollment-and- unemployment/](https://www.forbes.com/sites/richardvedder/2018/07/05/academic-deserted-villages/#90d18565121b)

University of Central Oklahoma, Office of Institutional Research <http://sites.uco.edu/academic-affairs/ir/>

United States Census Bureau (2018). *2017 ACS PUMS DATA DICTIONARY.* Retrieved from <https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2017.pdf>?

University of Central Oklahoma (2008-2017), Office of Institutional Research. *Factbook 2008-2017.* Retrieved from <http://sites.uco.edu/academic-affairs/ir/factbook.asp>

# **Appendix**

**Python code**

1. **Import the two dataframes**

import pandas as pd

data1=pd.read\_csv('D:/data/PROJECT DATA/2008 to 2012/ss12pok.csv')

pd.set\_option('display.max\_column', 200)

data1()

import pandas as pd

data2=pd.read\_csv('D:/data/PROJECT DATA/2013 to 2017/psam\_p40.csv')

pd.set\_option('display.max\_column', 200)

data2.head()

1. **Rename the selected columns**

data1.rename(columns={"SERIALNO":"year","SCHG":"enrollment","SCHL":"Educational attainment","ESR":"Employment Status","SCH":"College Type","SEX":"gender","WAGP":"wage/salary","AGEP":"age"},inplace=True)

data1.head()

data2.rename(columns={"SERIALNO":"year","SCHG":"enrollment","SCHL":"Educational attainment","ESR":"Employment Status","SCH":"College Type", "SEX":"gender","WAGP":"wage/salary","AGEP":"age"},inplace=True)

data2.head()

1. **Join the two dataframes**

a = data1[['year','enrollment','Educational attainment','Employment Status','College Type','gender','wage/salary','age']]

b = data2[['year','enrollment','Educational attainment','Employment Status','College Type','gender','wage/salary','age' ]]

c=pd.concat([a,b])

c

1. **Prepare year column**

c["year"]=(c["year"].round(-9))/1000000000

c["year"]=c["year"].astype(int)

c.head()

1. **Prepare enrollment column**

c["enrollment"].fillna(0,inplace=True)

c.head()

def change\_col(x):

for i in [x]:

if i ==15:

return 'undergraduate'

elif i==16:

return 'graduate'

elif i==0:

return 'not enrolled'

else:

return 'enrolled but not in college'

c["enrollment"]=c['enrollment'].apply(change\_col)

c.head()

c.tail()

*Replacing with numerical values:*

replace = {'undergraduate': '1',

'graduate': '2',

'not enrolled': '3',

'enrolled but not in college': '4'}

c['em']=c['enrollment'].replace(replace)

c['em']=c['em'].astype('int')

c.head()

c.tail()

cpivot = c.pivot\_table(index ='year', columns ='enrollment',values ='em', aggfunc='count', dropna =True, fill\_value=True, margins=True)

cpivot.iloc[:,0:]\*100

1. **Prepare Educational attainment column**

def change\_column(x):

for i in [x]:

if i ==16:

return 'high sch/GED'

elif i==17:

return 'high sch/GED'

elif i==18:

return 'high sch/GED'

elif i==19:

return 'high sch/GED'

elif i==20:

return 'college degree'

elif i==21:

return 'college degree'

elif i==22:

return 'college degree'

elif i==23:

return 'college degree'

elif i==24:

return 'college degree'

elif i==0:

return 'no formal education'

elif i==1:

return 'no formal education'

else:

return 'no diploma'

c["Educational attainment"]=c['Educational attainment'].apply(change\_column)

c.head()

c.tail()

*Replacing with numerical values:*

replace = {'high sch/GED': '1',

'college degree': '2',

'no formal education': '3',

'no diploma': '4'}

c['ae']=c['Educational attainment'].replace(replace)

c.head()

c.tail()

eapivot=c.pivot\_table(index ='year', columns ='Educational attainment',values ='ae', aggfunc='count', dropna =True, fill\_value=True, margins=True)

eapivot.iloc[:,0:]\*100

1. **Prepare Employment Status column**

def change(x):

for i in [x]:

if i ==0:

return 'less than 16 years old'

elif i==3:

return 'unemployed'

elif i==6:

return 'not in the labor force'

else:

return 'employed'

c["Employment Status"]=c['Employment Status'].apply(change)

c.head()

c.tail()

*Replacing with numerical values:*

replace = {'less than 16 years old': '1',

'unemployed': '2',

'not in the labor force': '3',

'employed': '4'}

c['es']=c['Employment Status'].replace(replace).astype(int)

c.head()

c.tail()

espivot = c.pivot\_table(index ='year', columns ='Employment Status',values ='es', aggfunc='count', dropna =True, fill\_value=True, margins=True)

espivot.iloc[:,0:]\*100

1. **Prepare College Type column**

c["College Type"].fillna(0,inplace=True)

c["College Type"]=c["College Type"].astype(int)

def col\_type(x):

for i in [x]:

if i ==3:

return 'private college'

elif i==2:

return 'public college'

elif i==0:

return 'less than 3 years old'

else:

return 'not attended college in 3 months'

c["CT"]=c["College Type"].apply(col\_type)

c.head()

c.tail()

ctpivot=c.pivot\_table(index ='year', columns ='CT',values ='College Type', aggfunc='count', dropna =True, fill\_value=True, margins=True)

ctpivot.iloc[:,0:]\*100

1. **The other variables are ready for use. We just have to make sure they are all integers**

c["gender"]=c["gender"].astype(int) # 1 = male and 2 = female

def sex(x):

for i in [x]:

if i==1:

return 'male'

else:

return 'female'

c["sex"]=c["gender"].apply(sex)

c

c["wage/salary"].fillna(0,inplace=True)

c["wage/salary"]=c["wage/salary"].astype(int)

c["age"].fillna(0,inplace=True)

c["age"]=c["age"].astype(int)

c.head()

c.tail()

**Data Visualization and Exploration**

1. **Import libraries**

import seaborn as sns

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

1. **Filter enrollment coloumn**

filt1=["undergraduate","graduate"]

colstudent=c[c["enrollment"].isin(filt1)]

colstudent.head()

filt1=["undergraduate"]

undergraduate=c[c["enrollment"].isin(filt1)]

undergraduate.head()

filt1=["graduate"]

graduate=c[c["enrollment"].isin(filt1)]

graduate.head()

1. **Explore and Visualize enrollment column**

cem = colstudent.pivot\_table(index ='year', columns ='enrollment',values ='em', aggfunc='count').astype(int)

%matplotlib inline

cem.plot(figsize=(16,4))

#*undergraduates only*

emug=undergraduate.pivot\_table(index ='year', columns ='enrollment',values ='em', aggfunc='count').astype(int)

%matplotlib inline

emg.plot(figsize=(16,4))

*# enrollment by gender*

boy = colstudent.pivot\_table(index ='year', columns ='sex',values ='em', aggfunc='count').astype(int)

%matplotlib inline

boy.plot(figsize=(16,4))

c.groupby(['year','enrollment','gender']).agg({'age': 'mean'}).tail(50)

c.groupby(['year','enrollment','gender']).agg({'em': 'count'}).tail(50)

colstudent.pivot\_table(index ='year', columns ='enrollment',values ='em', aggfunc='count', dropna =True, fill\_value=True, margins=True)

# *enrollment by gender*

colstudent.groupby(['year','enrollment','gender']).agg({'em': 'count'})

1. **Visualize the 'age' column**

*# Age distribution in the population*

sns.distplot(c['age']);

c['age'].describe()

*# average age for each enrollment category*

csage = c.pivot\_table(index ='year', columns ='enrollment',values ='age', aggfunc='mean').astype(int)

csage

*# Age distribution of college students in the population - both undergraduates and graduates*

sns.distplot(colstudent['age']);

colstudent['age'].describe()

c.groupby(['gender']).agg({'age': ['mean','min', 'max']}).astype(int)

*#indepth look at age distribution #average age by enrollment catgory*

colstudent.groupby(['year','enrollment']).agg({'age': ['mean','min', 'max']}).astype(int)

*#average age by enrollment catgory and gender*

colstudent.groupby(['year','enrollment','gender']).agg({'age': ['mean','min', 'max']}).astype(int)

*# Age distribution of undergraduates*

sns.distplot(undergraduate['age']);

undergraduate['age'].describe()

*# Age distribution of graduates*

sns.distplot(graduate['age']);

graduate['age'].describe()

1. **Visualize the 'Educational attainment' column**

ea = c.pivot\_table(index ='year', columns ='Educational attainment',values ='ae', aggfunc='count').astype(int)

%matplotlib inline

ea.plot(figsize=(16,4))

filt1=["enrolled but not in college"]

highsch=c[c["enrollment"].isin(filt1)]

highsch.head()

highsch.groupby(['enrollment', 'sex']).agg({'em': 'count'})

fig, ax = plt.subplots(figsize=(18, 5))

sns.countplot(x='Educational attainment', data=c, hue='sex', ax=ax);

c.groupby(['Educational attainment', 'sex']).agg({'ae': 'count'})

c.groupby(['Educational attainment', 'sex']).agg({'ae': 'count','wage/salary':'mean'}). astype(int)

c.groupby(['year','Educational attainment','sex']).agg({'ae': 'count', 'wage/salary':'mean'})

c.groupby(['Educational attainment','gender']).agg({'ae': 'count'})

*#make pie chart*

c.groupby(['gender']).agg({'gender': 'count'})

*#5. analyze 'wage/salary' column*

sns.distplot(c['wage/salary']);

c['wage/salary'].describe()

*#analyze wage/salary by gender*

c.groupby(['gender']).agg({'wage/salary': ['mean','min', 'max', 'std','size']})

sns.boxplot(x='salary', data=emp);

sns.boxplot(x='wage/salary', data=c);

c.groupby(['year']).agg({'wage/salary': ['mean', 'max', 'std','size']}).astype(int)

yy = c.pivot\_table(index ='year', values ='wage/salary', aggfunc='mean').astype(int)

%matplotlib inline

yy.plot(figsize=(16,4))

c.groupby(['sex']).agg({'wage/salary': 'mean'}).astype(int)

c.groupby(['sex']).agg({'wage/salary': 'mean'}).plot(kind='bar')

c.groupby(['year','gender']).agg({'wage/salary': 'mean'})

wy = c.pivot\_table(index ='year', columns ='sex',values ='wage/salary', aggfunc='mean').astype(int)

%matplotlib inline

wy.plot(figsize=(16,4))

c['Employment Status'].value\_counts().plot(kind='pie',figsize=(8,8), autopct='%1.1f%%')

c.groupby(['Employment Status','gender']).agg({'es': 'count'})

c.groupby(['Educational attainment','gender']).agg({'ae':'count'})

fig, ax = plt.subplots(figsize=(18, 5))

sns.countplot(x='Employment Status', data=c, hue='sex', ax=ax);

c.groupby(['year','Employment Status']).agg({'es':['count']})

c.groupby(['Educational attainment','Employment Status','gender']).agg({'ae': ['count'], 'es':['count']})

colstudent.groupby(['year','CT']).agg({'College Type': ['count']})

1. **Analyze and visualize 'employment status' column**

c.groupby(['Employment Status','gender']).agg({'es': 'count'})

c.groupby(['year','Employment Status','gender']).agg({'es': 'count'})

colstudent.groupby(['year','Employment Status','enrollment','gender']).agg({'es': 'count'})

**Regression and Statistic Test**

1. **Time series analysis**

yearly=c.copy()

yearly["year"].dtype

pd.to\_datetime(yearly.year, format='%Y').head()

yearly.head(2)

yearly["year"].unique()

yearly["enrollment"].unique()

yearly["year"].tail()

yearly["year"] =yearly["year"].apply(lambda x: pd.to\_datetime(str(x), format='%Y'))

import seaborn as sns

sns.set(style="darkgrid")

import matplotlib.pyplot as plt

%matplotlib inline

yearly.resample('Y',on='year').agg({'enrollment': 'sum'}).plot(kind="line",figsize=(20,10))

plt.xticks(size = 15)

plt.yticks(size = 15)

plt.title('Year-wise total enrollment', size = 15)

plt.xlabel('Year', size = 15)

plt.ylabel('Enrollment',size=15)

plt.legend(prop={'size':15})

histo=yearly.resample('Y',on='year').agg({'enrollment': 'sum'})

histo

yearly['year'] = yearly['year'].dt.year

histo=histo.reset\_index()

histo

sns.catplot(x="year", y="enrollment",hue="gender",data=yearly,kind='bar', height=5, aspect=2);

c["enrollment"].unique()

c.tail()

1. **Ensure all data for the later use are integers**

c["gender"]=c["gender"].astype(int) # 1 = male and 2 = female

c["wage/salary"].fillna(0,inplace=True)

c["wage/salary"]=c["wage/salary"].astype(int)

c["age"].fillna(0,inplace=True)

c["age"]=c["age"].astype(int)

c.head()

filt1=["undergraduate","graduate"]

filt2=["private college","public college"]

c2=c[c["enrollment"].isin(filt1)]

c3=c[c["college\_type\_new"].isin(filt2)]

cnew=pd.concat([c2,c3])

cnew.head(5)

**Creating dummy variables and merging them with the original data frame:**

c\_year=pd.get\_dummies(c["year"],prefix="c",drop\_first=False,dtype="int")

c2=pd.concat([c, c\_year], axis=1)

c2["Employment Status"].fillna(0,inplace=True)

c2["Employment Status"]=c2["Employment Status"].astype(int)

def empl(z):

for i in [z]:

if ((i ==1) | (i==2) |(i ==4) | (i==5)):

return 1

else:

return 0

c2["cat\_Employment"]=c2["Employment Status"].apply(empl)

c\_Educational\_attainment=pd.get\_dummies(c2["Educational attainment"],prefix="c",drop\_first=False,dtype="int")

c3=pd.concat([c2,c\_Educational\_attainment], axis=1)

#School type

# b .N/A (less than 3 years old)

# 1 .No, has not attended in the last 3 months

# 2 .Yes, public school or public college

# 3 .Yes, private school or college or home school

c\_school\_type=pd.get\_dummies(c3["College Type"],prefix="school",drop\_first=False,dtype="int")

c4=pd.concat([c3,c\_school\_type], axis=1)

**Dropping the unnecessary columns:**

c4.drop(columns=["year","Educational attainment",\

"Employment

Status","College Type","gender","Attain\_new","college\_type\_new"],inplace=True)

**Finding the Correlations:**

c4\_corr=c4[["wage/salary","age","c\_2008","c\_2009","c\_2010","c\_2011","c\_2012","c\_2013","c\_2014",\

"c\_2015","c\_2016","c\_2017","cat\_Employment","c\_college degree","c\_high sch/GED",\

"c\_no diploma","c\_no formal education","school\_0","school\_1","school\_2","school\_3","sex"]]

c4\_corr.corr()

**Creating the heatmap:**

import numpy as np

from sklearn import preprocessing

import matplotlib.pyplot as plt

plt.rc("font", size=14)

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

import statsmodels.formula.api as smf

import seaborn as sns

import matplotlib.pyplot as plt

plt.figure(figsize=(24,14))

sns.heatmap(c4\_corr.corr(),annot=True,cmap='coolwarm')

**Renaming few columns for regression:**

c4=c4.rename(columns={"wage/salary":'salary',"c\_high sch/GED":'c\_highschool',"c\_no diploma":'c\_nodiploma',"c\_no formal education":'c\_noedu'})

**Logistic regression using stats models:**

logfit=smf.logit('enrollment ~ salary + age + C(c\_2009)\

+C(c\_2010)+C(c\_2011) + C(c\_2012) + C(c\_2013) \

+ C(c\_2014) + C(c\_2015) + C(c\_2016) \

+ C(c\_2017)+ C(cat\_Employment) + C(c\_highschool) + \

C(c\_nodiploma) + C(c\_noedu) + C(sex) ',data=c4).fit()

logfit.summary()

logfit2=smf.logit('enrollment ~ age + C(c\_2009)+ +C(c\_2010)+C(c\_2011) + C(c\_2012) + C(c\_2013) + C(c\_2014) + C(c\_2015) + C(c\_2016) + C(c\_2017)+ C(cat\_Employment) + C(c\_highschool) + C(c\_nodiploma) + C(sex) ',data=c4).fit()

logfit2.summary()

**Finding odds ratios:**

model\_odds = pd.DataFrame(np.exp(logfit2.params), columns= ['OR']).round(2)

model\_odds['z-value']= logfit2.pvalues

model\_odds[['2.5%', '97.5%']] = np.exp(logfit2.conf\_int())

model\_odds['OR'] = model\_odds['OR'].apply(lambda x: '{:.2f}'.format(x))

model\_odds['z-value'] = model\_odds['z-value'].apply(lambda x: '{:.2f}'.format(x))

model\_odds['2.5%'] = model\_odds['2.5%'].apply(lambda x: '{:.2f}'.format(x))

model\_odds['97.5%'] = model\_odds['97.5%'].apply(lambda x: '{:.2f}'.format(x))

model\_odds

**Using Sci-kit learn package for accuracy and ROC curve:**

X=c4[["age","c\_2009","c\_2010","c\_2011","c\_2012","c\_2013","c\_2014","c\_2015","c\_2016","c\_2017","cat\_Employment","c\_highschool","c\_nodiploma","sex"]]

y=c4["enrollment"]

X.fillna(0,inplace=True)

y.fillna(0,inplace=True)

from sklearn.linear\_model import LogisticRegression

from imblearn.over\_sampling import SMOTE

os = SMOTE(random\_state=0)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=0)

columns = X\_train.columns

os\_data\_X,os\_data\_y=os.fit\_sample(X\_train, y\_train)

os\_data\_X = pd.DataFrame(data=os\_data\_X,columns=columns )

os\_data\_y= pd.DataFrame(data=os\_data\_y,columns=['y'])

logreg = LogisticRegression()

logreg.fit(X\_train,y\_train)

classifier=LogisticRegression(random\_state=0)

classifier.fit(X\_train,y\_train)

y\_pred=classifier.predict(X\_test)

y\_pred

from sklearn.metrics import accuracy\_score

accuracy\_score(y\_test,y\_pred)

**ROC curve:**

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve

logit\_roc\_auc = roc\_auc\_score(y\_test, logreg.predict(X\_test))

fpr, tpr, thresholds = roc\_curve(y\_test, logreg.predict\_proba(X\_test)[:,1])

plt.figure()

plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit\_roc\_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

plt.savefig('Log\_ROC')

plt.show()

**Time series Analysis:**

(Copying the same data for time series analysis)

yearly=c.copy()

yearly["year"].dtype

pd.to\_datetime(yearly.year, format='%Y').head()

yearly["year"] =yearly["year"].apply(lambda x: pd.to\_datetime(str(x), format='%Y'))

import seaborn as sns

sns.set(style="darkgrid")

import matplotlib.pyplot as plt

%matplotlib inline

yearly.resample('Y',on='year').agg({'enrollment': 'sum'}).plot(kind="line",figsize=(20,10))

plt.xticks(size = 15)

plt.yticks(size = 15)

plt.title('Year-wise total enrollment', size = 15)

plt.xlabel('Year', size = 15)

plt.ylabel('Enrollment',size=15)

plt.legend(prop={'size':15})

enroll\_yearly=yearly.resample('Y',on='year').agg({'enrollment': 'sum'})

yearly['year'] = yearly['year'].dt.year

histo=histo.reset\_index()