Statistical Analysis:
Economic Values of College Majors

MSBA 320, Advanced Statistical Analysis with Python and R

Professor Dr. Siamak Zadeh

Attempted By

Digvijay Bisla, Student ID: 0603803

Table of Content

Introduction

In today's contemporary world the purpose of going to college should not just be exclusively gaining financial value, developing skillsets, or just making money in the future. In it lie the impalpable benefits that help the students reach their highest potential. Though when the debts rises on these very students and the rate of nonemployment dips, they and their families are faced with the question that the college education they are planning will be worthwhile of the very investment they are putting in. Even though the perception of students might still be in mist, we as a community certainly see the great advantages of students who graduate and the even greater power that these colleges hold who cultivate these graduates. My dataset talks about the economic values of college majors. I tried finding the relationships between all the columns in college graduation including *Major Types*, *Employment category*, *Wages* etc. The research is going to get us acquainted with the most profitable portfolios to be a part of and that their exist.

Data Collection

My dataset has been selected from the webpage FiveThirtyEight.com(The Economic Guide to Picking a College Major), This shows the median salary for multiple Major degrees. The CSV file used in my dataset includes data from a Gitrepository which has a focus on current graduates as it is also having gender specific data.

Data Analysis

I commenced my research with the purpose of getting a good understanding of my dataset by evaluation of the data frame for data types, columns, and structure. So, I took a look at the dataset by displaying the first two observations in it as could be seen in the figure 1. Further downrange I decided to drop the columns that I thought would not be used as much for example *unemployment_rate, Median, P25th*. To evaluate the dataframe further I printed the datatypes by displaying the columns.

```
In [3]: #taking a look at our data set
        college_Majors.head(2)
```

Out[3]:

	Rank	Major_code	Major	Total	Men	Women	Major_category	ShareWomen	Sample_size	Employed	...	Part_time	Full_time_year_round	Unemployed	Unemployment_rate	Median	P25th	P75th	College_jobs	Non_college_jobs	Low_wage_jobs
0	1	2419	PETROLEUM ENGINEERING	2339.0	2057.0	282.0	Engineering	0.1206	36	1976	...	270	1207	37	0.0184	110000	95000	125000	1534	364	193
1	2	2416	MINING AND MINERAL ENGINEERING	756.0	679.0	77.0	Engineering	0.1019	7	640	...	170	388	85	0.1172	75000	55000	90000	350	257	50

2 rows × 21 columns

Figure 1. First 2 Rows

Data Cleaning

Further I decided to check if my variables had any null values and it came out to be that 3 of my variables did have null values. Further downrange I replaced all nulls with zeroes. Since I needed to have a singular data type, so I checked the datatypes of my data. It came out to be that 3 data types which were printed as *float64* and needed to be changed to *Int64* datatype. After summarizing the dataset our observation count came out to be as follows: -

```
[173 rows x 16 columns]>
```

Figure 2. Shape of our dataset

Data Visualization

I decided to use and determine the positive relationships using Pearson correlation. Following relationships, we were able to identify as could be seen in figure 3 below: -

- college_jobs vs men (0.56)
- college_jobs vs women (0.85)
- full_time_year_round vs men (0.89)
- full_time_year_round vs women (0.91)
- college_jobs vs Employed (0.80)
- college_jobs vs unemployed (0.71)
- full_time_year_round vs Employed (0.99)

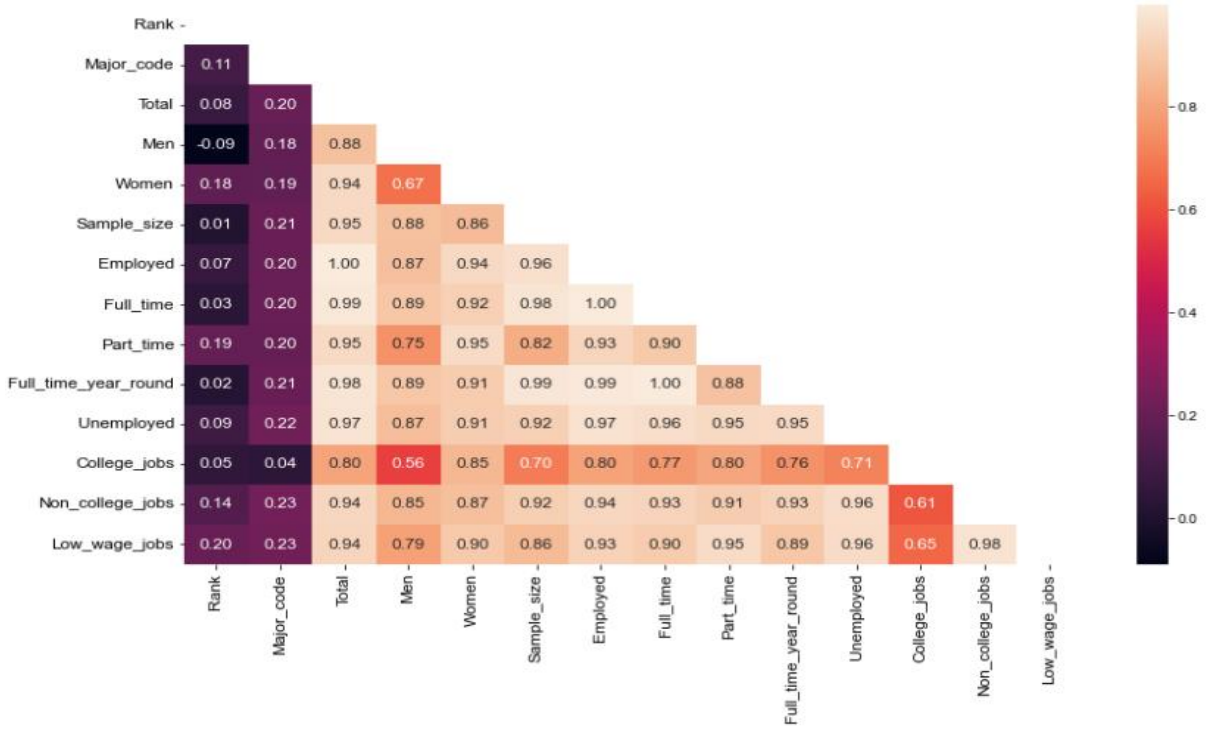

Figure 3. *Heat Map*

Further I decided plotting a histograms to determine the best pairs. The columns are as follows:-

number_features=[*'Major_code','Total','Men', 'Women', 'Employed','Full_time', 'Part_time',*

'Full_time_year_round', 'Unemployed','College_jobs', 'Non_college_jobs','Low_wage_jobs']

Histograms comprising of only numeric plots of the distribution were printed and the results may be seen in the figure

- Looking at the histogram with variable *Men* it could be observed that it is fairly continuous, the maximum point lies within 0 and 1 with the maximum count above 80. In the lowest over 0 is 175000.
- Looking at the histogram with variable *Employed* it could be observed that it is fairly continuous, the maximum point lies within 0 and 1 with the maximum count at around 80. In the lowest over 0 is 300000.
- Looking at the histogram with variable *Women* it could be observed that it is fairly continuous, the maximum point lies within 0 and 1 with the maximum count above 80. In the lowest over 0 is 300000.
- Looking at the histogram with variable *Fulltime* it could be observed that it is fairly continuous, the maximum point lies within 0 and 1 with the maximum count at around 80. In the lowest over 0 is 250000.
- Looking at the histogram with variable *Parttime* it could be observed that it is fairly continuous, the maximum point lies within 0 and 1 with the maximum count above 80. In the lowest over 0 is 120000.

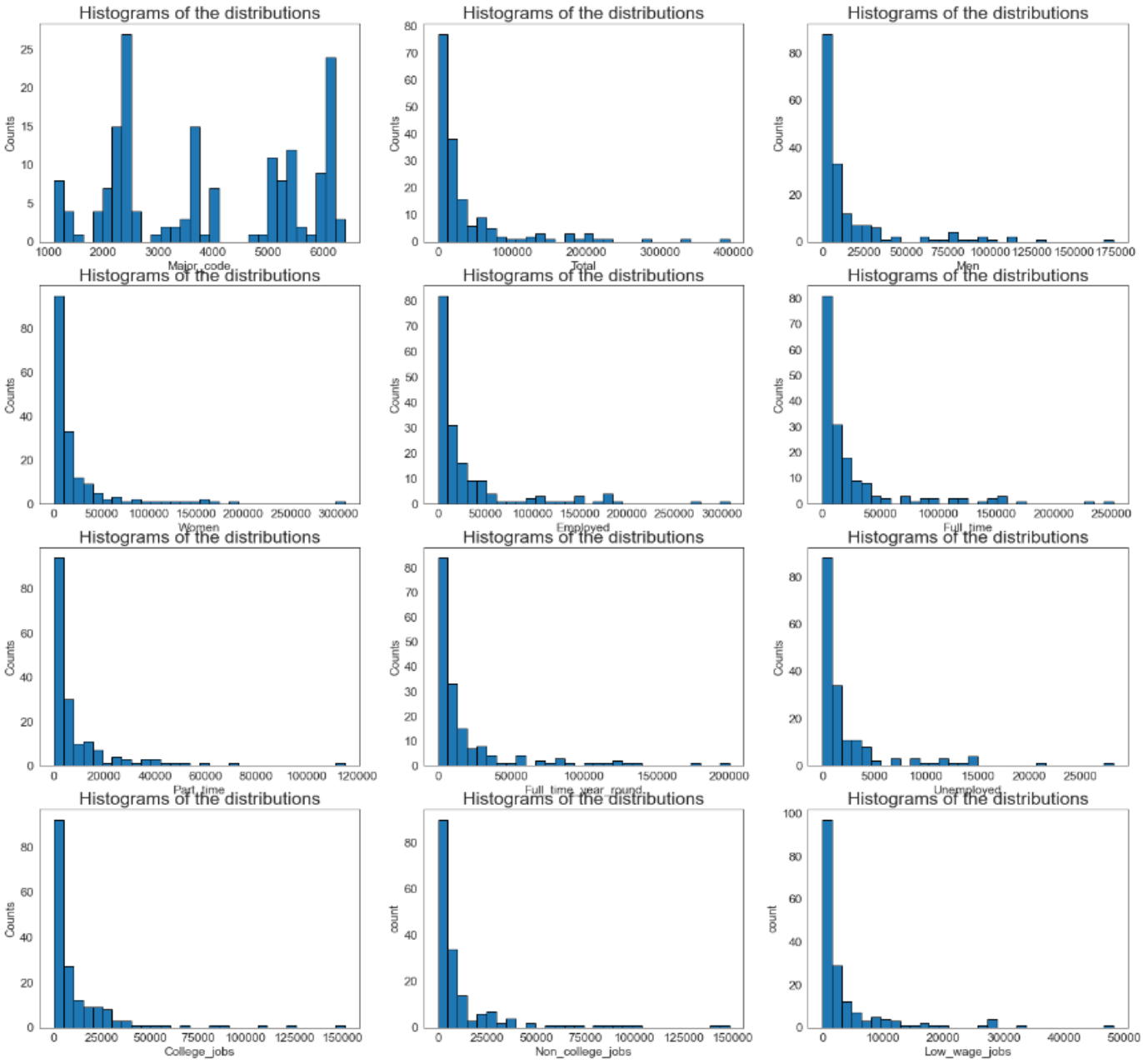

Figure 4. Distribution ***Histograms***

Since the research was going to get us acquainted with the most profitable portfolios to be a part of and that their exist. I planned on visualizing the categorical data posing my **first question** of top 10 Majors categories with highest median salary. I found very indulging and interesting results as could be seen in figure 5 that the best Major degree was from Engineering and came out to be *Mechanical Engineering* had the highest median salary and it was then followed by *Electrical engineering* appearing on the second, the third one was *Chemical engineering* followed by *Actuarial science* and *Nuclear Engineering*.

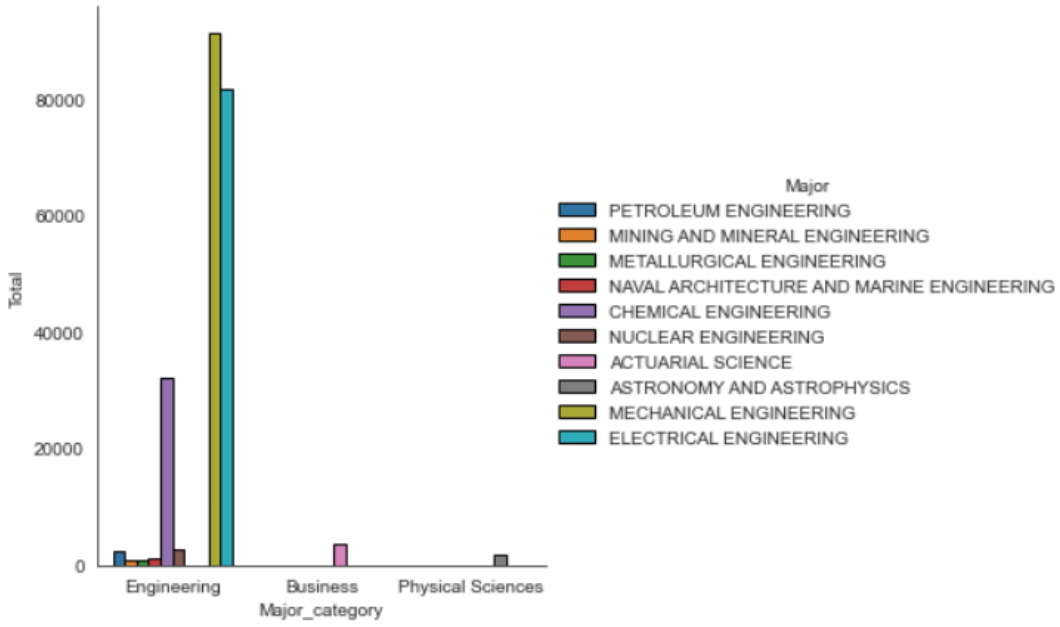

Figure 5. visualizing the top 10 major categories

Moving further in my research the **second question** that popped in my mind was to make a count of Major categories their exist in my data and the count of majors there are in each one of them. So, I decided to create a Bar plot of Major Categories as could be seen in Figure 6 and I came up with the following observations. *Engineering* Domain of Majors have the highest major degrees count in it followed by *Education* and *Humanities Liberal Arts*. The lowest count came out to be in *Communication & Journalism* as well as *Indisplinary*.

We all have a fair idea that what engineers do in the today's contemporary world and the essence they hold and play in our society. This is what lead the American Association of Engineering Societies(How to add this as reference) to select engineering as the stealth profession. Engineering is really the backbone of human civilization and same could be seen in my research that the Majors included in this domain have the highest value and this degree of being renowned has led them to expand their categorization as could be seen in below mentioned Figure 6.

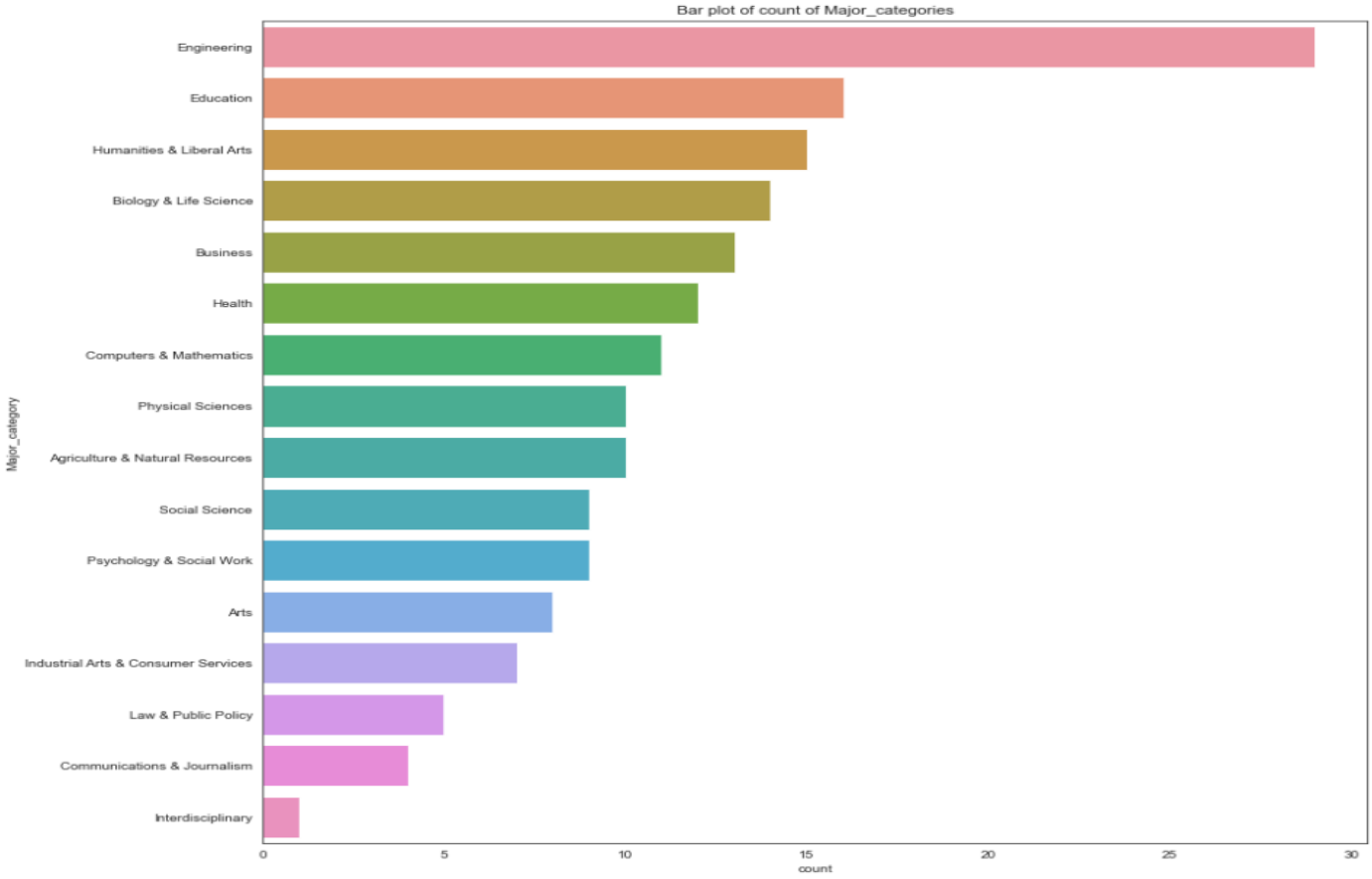

Figure 6. Bar plot of Majors count in Major_categories

Moving further in my research the **third question** I went on asking was how extensive the unemployment is ranging from Major to Major. So, I decided to make a bar plot as could be seen in figure 7. Here the line shows the error in the data. After receiving the results, I was not at state of aww at all as the highest rate of unemployment was observed in *communication and journalism*, *Business* and *social science*.

After taking majors like these it happens to be when the debts rises on these students and the rate of nonemployment dips, they and their families are faced with the question that the college education they are planning will be worthwhile of the very investment they are putting in.

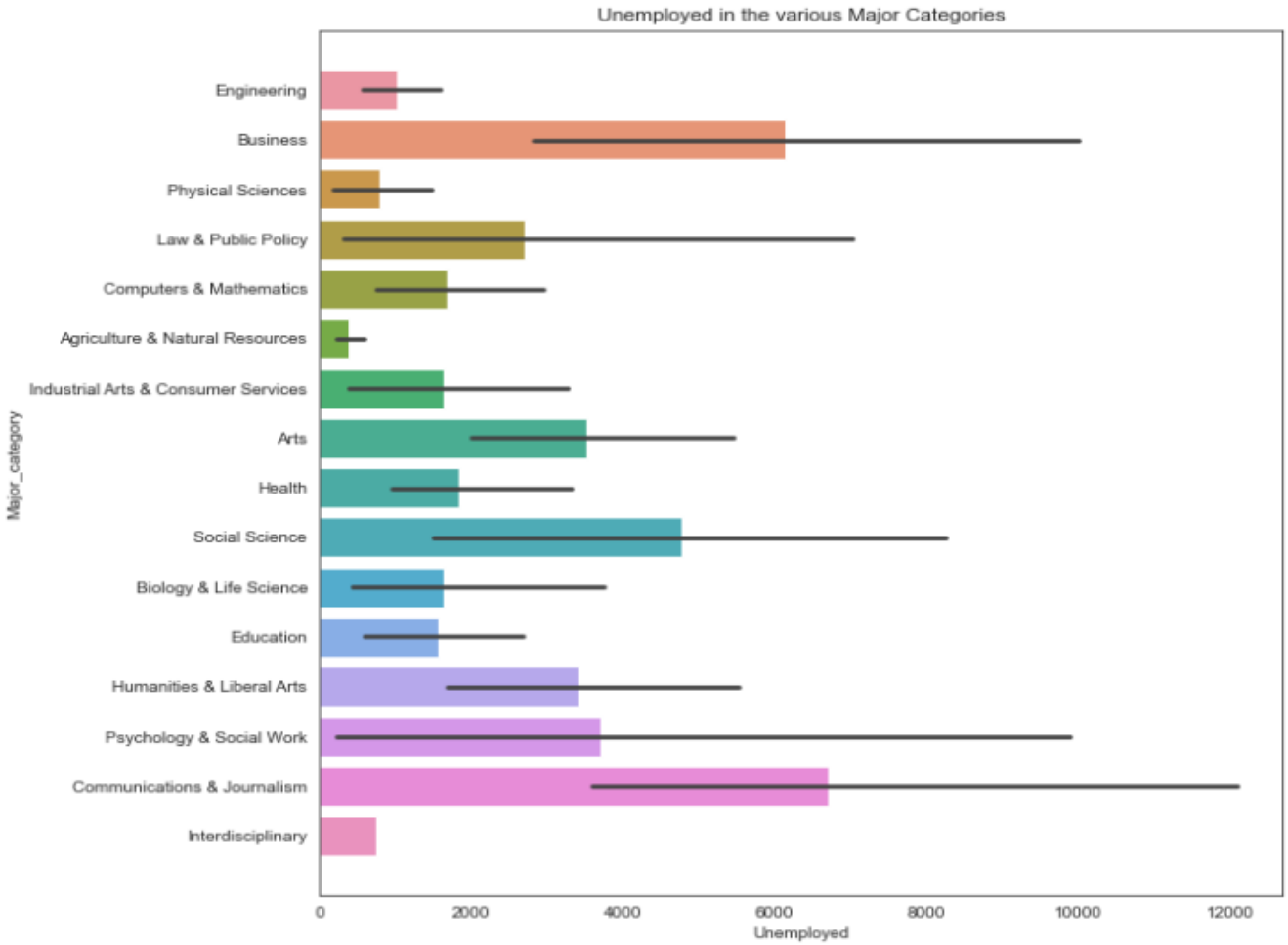

Figure 7. Bar plot for Unemployed in the various Major Categories

Moving further in my research the **fourth question** I wanted to solve was Which course do men take the highest. So, I decided to make a bar plot as could be seen in figure 8. It comes out to be that most preferred course by men was *Business* followed by *Communication and journalism* with least preferred being *indisciplenary* courses. If we consider how much monetary returns are their from Engineering degrees and other technical Majors I thought that these will be the most popular ones among men but the results came out to be totally unexpected. Degrees like *Communication and journalism* even after being low paying came out to be most popular ones. What I came out to believe that due to the difficulty of these technical Majors very few Male participants found them appealing. Visualization may be seen in the appended figure 8.

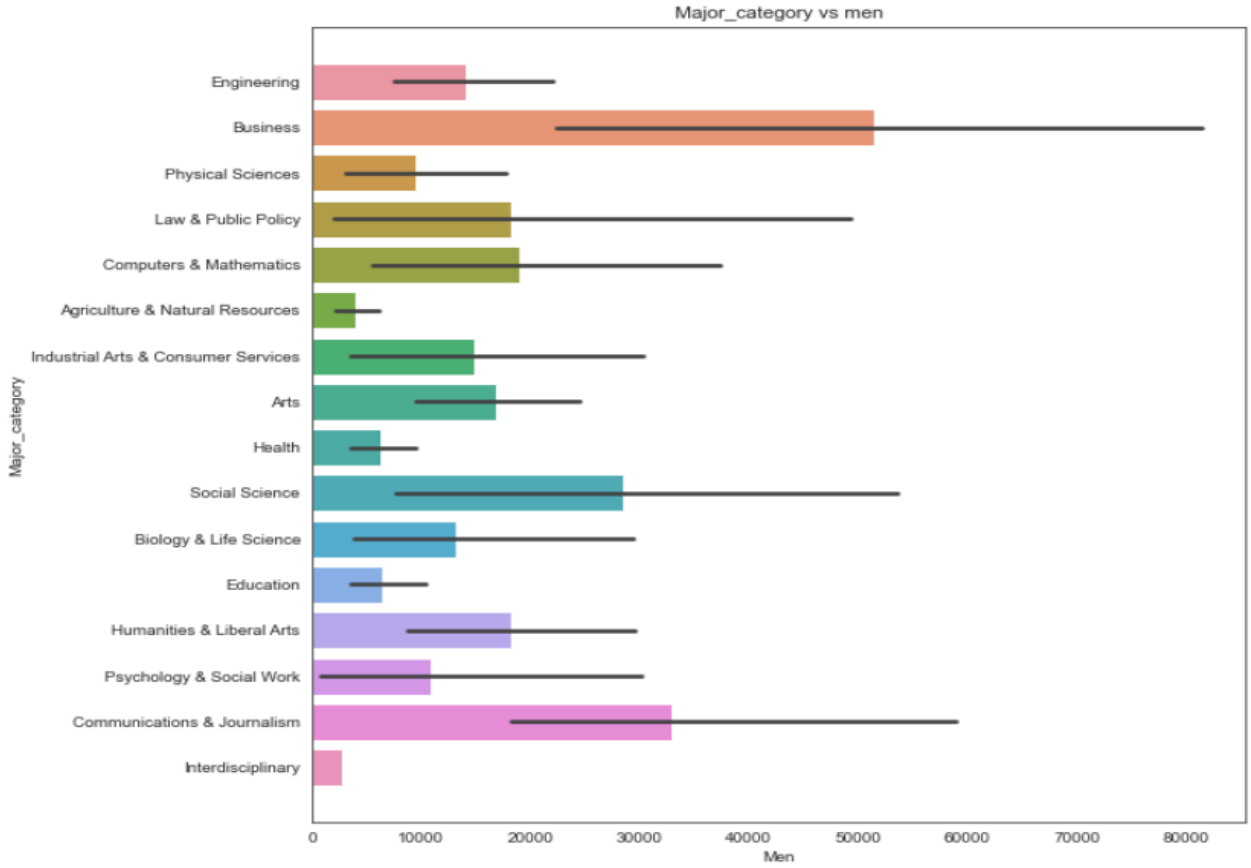

Figure 8. Bar plot for Major_category vs men

Moving further in my research the **fifth question** I wanted to solve was Which course do women take the highest. So, I decided to make a bar plot as could be seen in figure 9. It comes out to be that most preferred course by women was *Communication and journalism* Followed by *business* and *Psychology & Social work.* Compared to what we found in the Men's visualization women were found to be more dominant in the field of healthcare and Psychology. Though men continue to dominate the technical fields women found it least engrossing. I have developed a clear understanding through the below analysis that women prefer the business profession as much as men do.

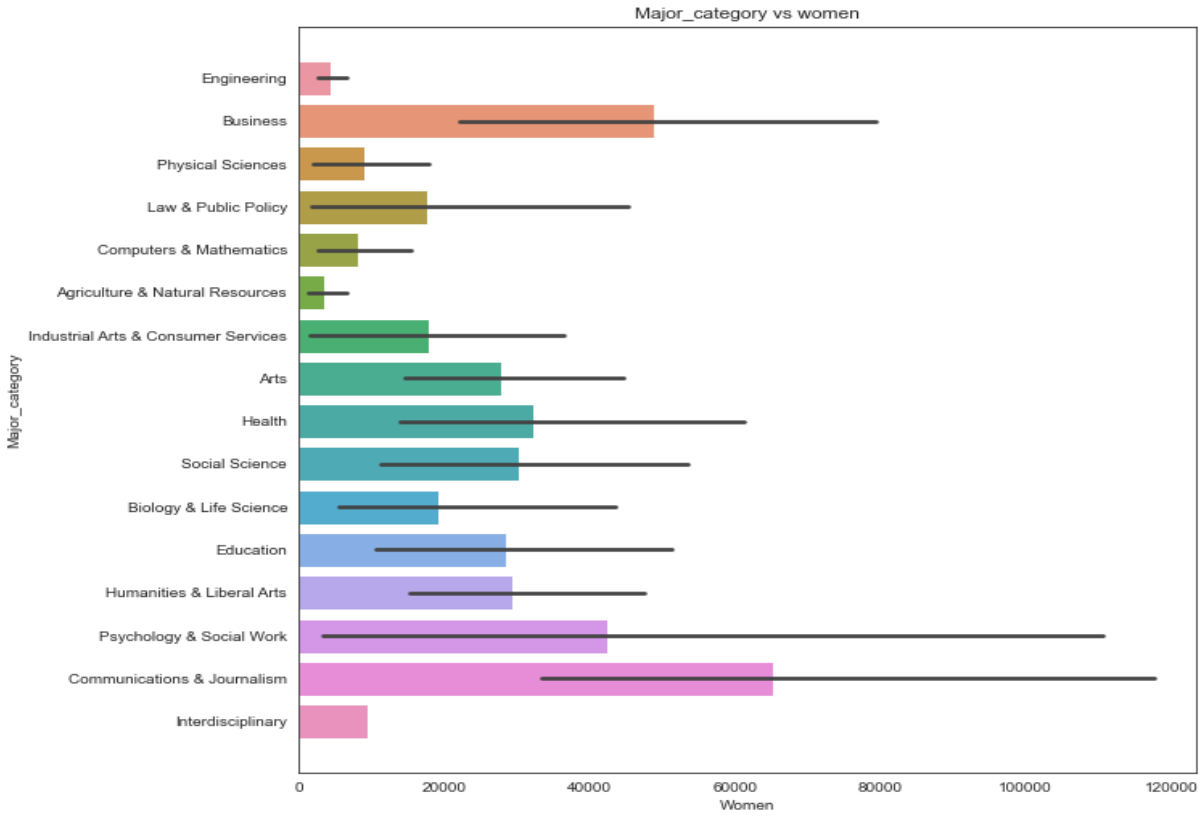

Figure 9. Bar plot for Major_category vs women

Regression Analysis

Moving further in my research I used pair plots to help me in identifying the perfect relationships i.e. finding the relationships between variables that show the influence of one variable over the other, especially on the target. So we ran univariate regression and also applied the line of fit to help us find the fit that is best as could be seen in figure 10

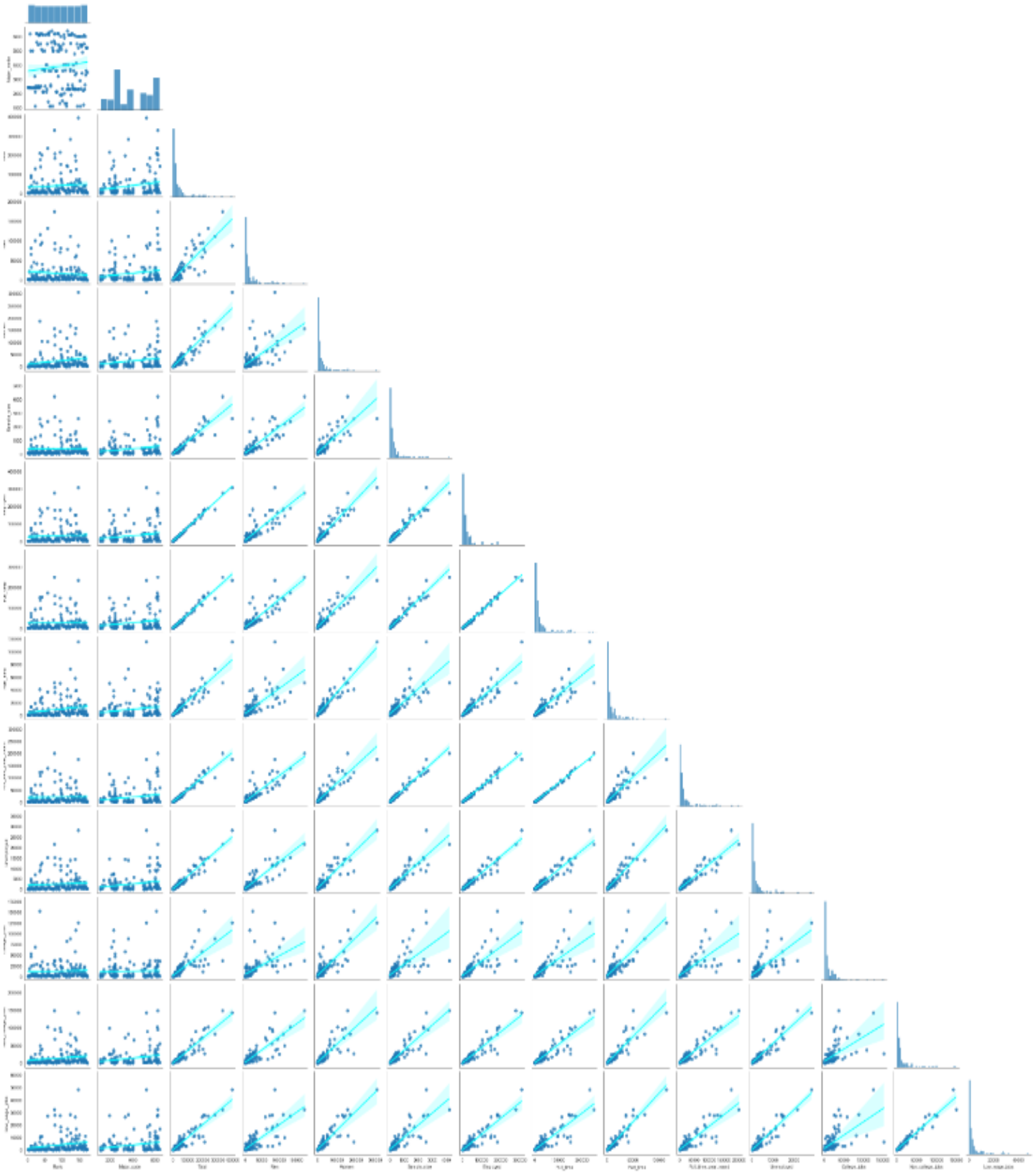

Figure 10. ***Pair Plot to check best fit***

Testing Model 1

Following was my summary and observation of the model testing for *Unemployed* vs *Men.* Same could be seen in below figure 11.

- Dependent Variable - *Men*
- Independent Variables - *Unemployed*

OLS Regression Results

Dep. Variable:	Men	**R-squared:**	0.756
Model:	OLS	**Adj. R-squared:**	0.755
Method:	Least Squares	**F-statistic:**	530.5
Date:	Tue, 13 Dec 2022	**Prob (F-statistic):**	2.69e-54
Time:	20:23:12	**Log-Likelihood:**	-1894.8
No. Observations:	173	**AIC:**	3794.
Df Residuals:	171	**BIC:**	3800.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2285.6927	1226.506	1.864	0.064	-135.348	4706.734
Unemployed	5.9351	0.258	23.033	0.000	5.426	6.444

Omnibus:	47.116	**Durbin-Watson:**	1.852
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	904.500
Skew:	0.230	**Prob(JB):**	3.89e-197
Kurtosis:	14.192	**Cond. No.**	5.52e+03

Figure 11. OLHS Regression result for model testing for Unemployed vs Men

In Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 5.52e+03. This might indicate that there are strong multicollinearity or other numerical problems.

My further interpretation came out to be that the value of correlation coefficient was 0.87. R-squared value came out to be 0.756, thus with use of the given predictors variability for our dependent variable for model 1 could be explained i.e. 75.6%. Further it could be seen in the above figure 11 that *Adj. R-squared* is lesser than *R-squared* which generally could be seen due to the

presence of an insignificant variable in the model. In addition to this they was less presence of standard error. Probability was observed to be 0.064 stating that if is greater than 0.05 then it is not statistically significant.

Moving further in my analysis I had to find the fit our model to determine if Unemployed has a very strong relationship with men. Same interpretation could be observed in figure 12 which states that weather the relationship is strong it's a mix match and the data is scattered meaning there is error in the data .

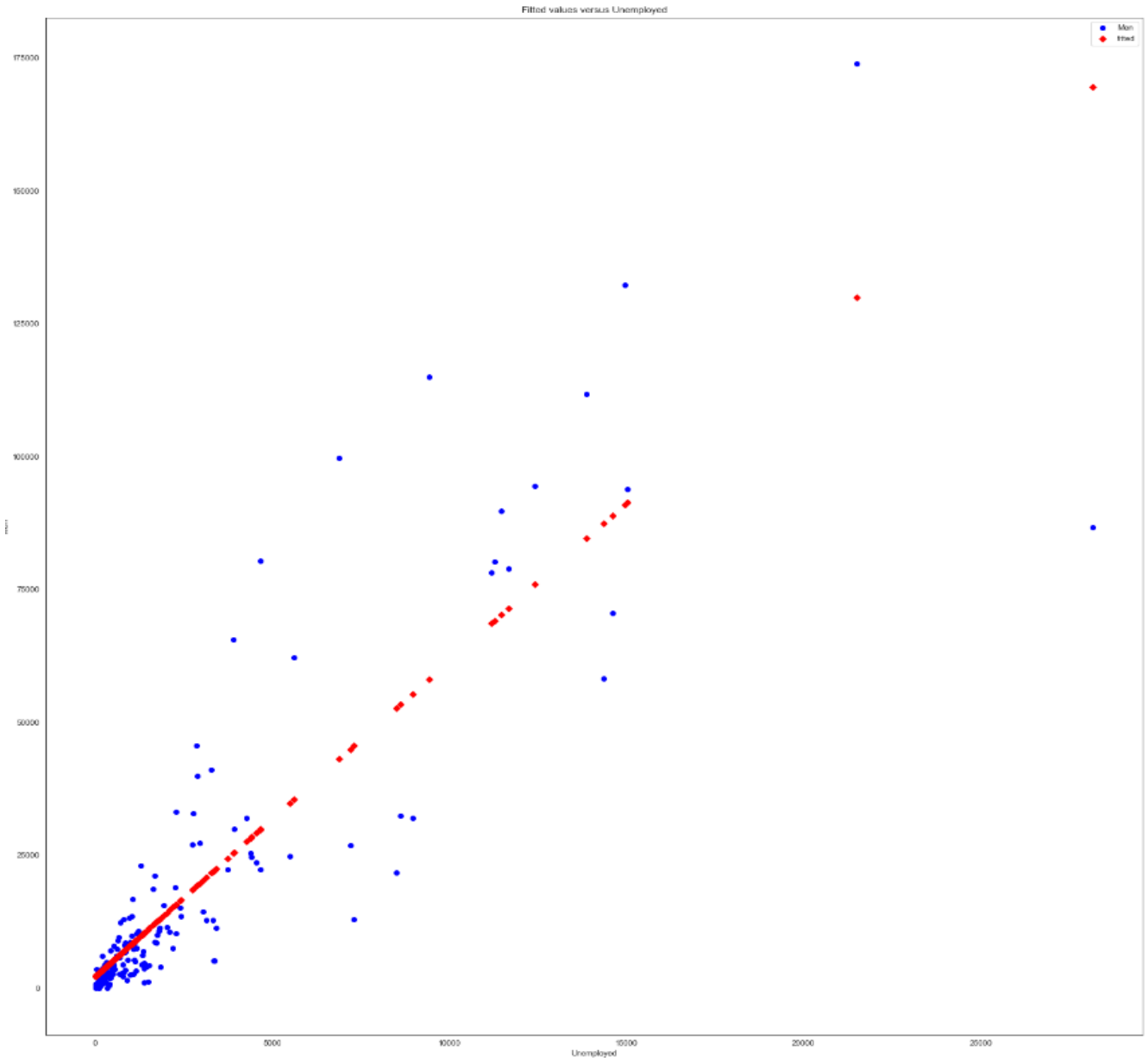

Figure 12. Plot to check fit our model

Testing Model 2

Following was my summary and observation of the model testing for *Unemployed* vs *Women*. Same could be seen in below figure 13.

- Dependent Variable - *Women*
- Independent Variables - *Unemployed*

Out[38]:

OLS Regression Results

Dep. Variable:	Women	**R-squared:**	0.831
Model:	OLS	**Adj. R-squared:**	0.830
Method:	Least Squares	**F-statistic:**	843.5
Date:	Thu, 15 Dec 2022	**Prob (F-statistic):**	5.16e-68
Time:	10:49:32	**Log-Likelihood:**	-1928.3
No. Observations:	173	**AIC:**	3861.
Df Residuals:	171	**BIC:**	3867.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	565.4105	1488.824	0.380	0.705	-2373.430	3504.251
Unemployed	9.0842	0.313	29.042	0.000	8.467	9.702

Omnibus:	122.080	**Durbin-Watson:**	1.999
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1948.868
Skew:	2.297	**Prob(JB):**	0.00
Kurtosis:	18.788	**Cond. No.**	5.52e+03

Figure 13. OLHS Regression result for model testing for Unemployed vs Women

In Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 5.52e+03. This might indicate that there are strong multicollinearity or other numerical problems.

My further interpretation came out to be that the value of correlation coefficient was 0.91. R-squared value came out to be 0.831, thus with use of the given predictors variability for our dependent variable for model 2 could be explained i.e. 83.1%. Further it could be seen in the above figure 13 that *Adj. R-squared* is lesser than *R-squared* which generally could be seen due to the presence of an insignificant variable in the model. In addition to this they was less presence of

standard error. Probability was observed to be 0.705 stating that if is greater than 0.05 then it is not statistically significant.

Moving further in my analysis I had to find the fit our model to determine if Unemployed has a very strong relationship with women. Same interpretation could be observed in figure 14 which states that weather the relationship is strong it's a mix match and the data is scattered meaning there is error in the data .

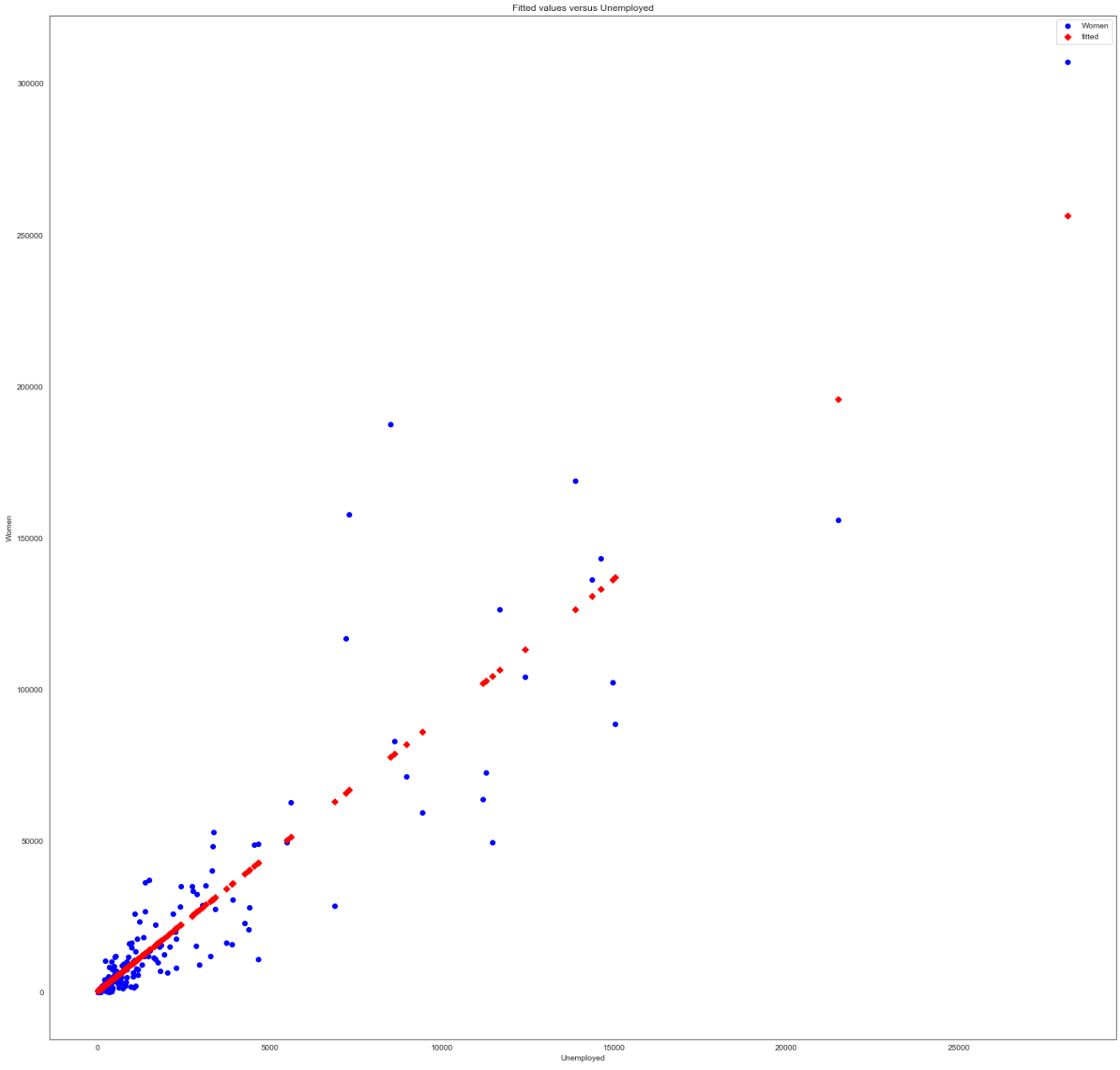

Figure 14. Plot to check fit our model

Conclusion

There existed high level of nervousness and apprehension every season for the students who planned to take on higher education. Questions like Majors they should pick to get the highest median salary or they competition their exists in these degrees. Which one should they take based of gender, pay, or just job security. With these objectives in mind and equipped with what I learned in class I wanted to find the solution for the same.

It came out to be posing my first question of top 10 Majors categories with highest median salary. I found very indulging and interesting results that the best Major degrees with highest median salary were from Engineering. Later in my research I came to a understanding that *Engineering* Domain of Majors have the highest major degrees count in it. Further I went on asking was how extensive the unemployment is ranging from Major to Major and observed nontechnical degrees like *communication & journalism and psychology* were at the top. Best monetary returns are there from Engineering degrees and other technical Majors I thought that these will be the most popular ones among men and women, but the results came out to be totally unexpected. Degrees like *Communication and journalism* even after being low paying came out to be most popular ones.

Engineering definitely came out to be the wild outlier, in terms of monetary benefits and both our variables of male and female behaved quiet similarly to it. A staggering seventeen out of the twenty majors with highest monetary payback were from the category of Engineering. That brings me to quoting the remark that every STEM category is not really same. Specially when I talk about the monetary payback the Science graduates collect a bigger paycheck then their counterparts with unavailability of full-time jobs in some categories also affecting the income.

Appendix

```
#importing the data set
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm

#importing the data
college_Majors = pd.read_csv("C:/Users/Aditya/Downloads/college_majors.csv")

#taking a look at our data set
college_Majors.head(2)

#displaying the columns
college_Majors.columns

#dropping the columns we will not need to use
college_Majors.drop(columns
=['P25th','P75th','Median','ShareWomen','Unemployment_rate'],inplace = True)

#lets check if we have any nulls in our data set
college_Majors.isnull().sum()

#replacing all the nulls with zeros
college_Majors.fillna(value=0, inplace = True)

#lets check if we have any zeros in our data set
college_Majors.isnull().sum()

#checking the data types of our data
college_Majors.dtypes

#convert all the floats to int
college_Majors.Total=college_Majors.Total.astype(int)
college_Majors.Men =college_Majors.Men.astype(int)
college_Majors.Women=college_Majors.Women.astype(int)

#let us get a summary of our data set
college_Majors.info

#we have two categorical columns Major and Major category
college_Majors.info()

#this gives us the shape of our data set which is 173 rows and 18 columns
college_Majors.shape

#we will get the information such as the mean, std, min and max plus the count of the
data
college_Majors.describe()

#check if our data set has is unique
college_Majors.Major_category.unique()
```

```
#we find the correlation of our data set using a heat map the by writing it in a
function
mask =np.zeros_like(college_Majors.corr())
major_triangle =np.triu_indices_from(mask)
mask[major_triangle] =True

plt.figure(figsize=(14,8))
sns.heatmap(college_Majors.corr(), mask=mask, annot=True,fmt='.2f',
annot_kws={"size":12})
sns.set_style('white')
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()
#relationsips we have identified
#college_jobs vs men
#college_jobs vs women
#full_time_year_round vs men
#full_time_year_round vs women
#college_jobs vs Employed
#college_jobs vs unemployed
#full_time_year_round vs Employed

#plotting a univariate analysis/ histograms to determin the best pairs
plt.rcParams['figure.figsize'] = (24,24)
#sub plots to be used
figures, axes=plt.subplots(nrows =4, ncols =3)
#columns used
number_features=['Major_code','Total','Men', 'Women', 'Employed','Full_time',
'Part_time',
       'Full_time_year_round', 'Unemployed','College_jobs',
'Non_college_jobs','Low_wage_jobs']
xaxes = number_features
yaxes
=['Counts','Counts','Counts','Counts','Counts','Counts','Counts','Counts','Counts','C
ounts','count','count']
#histograms of the distributions
axes =axes.ravel()

for j, a in enumerate(axes):
    a.hist(college_Majors[number_features[j]], bins = 30, edgecolor="black")
    a.set_title("Histograms of the distributions", fontsize=22)
    a.set_xlabel(xaxes[j], fontsize=15)
    a.set_ylabel(yaxes[j], fontsize=15)
    a.tick_params(axis='both', labelsize=15)

#visualizing the categorical data
#sns.catplot(data=college_Majors, x="Total", y="Major", hue="Major_category",
kind="bar")

college_Majors.Major_category.unique()

college_Majors.Major_category.value_counts()

college_Majors.Major.unique()

college_Majors.Major.value_counts()

categorica_data = college_Majors.head(10)
```

```
#visualizing the categorical data top 10 major categories
sns.catplot(data=categorica_data, x="Major_category", y="Total",edgecolor="black",
hue="Major", kind="bar")

plt.figure(figsize=(14,14))
plt.title("Bar plot of count of Major_categories")
sns.countplot(y = 'Major_category',
              data = college_Majors,
              order = college_Majors['Major_category'].value_counts().index)
plt.show()

#Unemployed vs Full time
#the line shows the error in the data
plt.figure(figsize=(10,10))
plt.title("Unemployed in the various Major Categories")
sns.barplot(data=college_Majors,x="Unemployed",y="Major_category")
plt.plot()

#Major_category vs men
plt.figure(figsize=(10,10))
plt.title("Major_category vs men")
sns.barplot(data=college_Majors,x="Men",y="Major_category")
plt.plot()
#Major_category vs women
plt.figure(figsize=(10,10))
plt.title("Major_category vs women")
sns.barplot(data=college_Majors,x="Women",y="Major_category")
plt.plot()

#Major_category vs Full_time
plt.figure(figsize=(10,10))
plt.title("Major_category vs Full_time")
sns.barplot(data=college_Majors,x="Full_time",y="Major_category")
plt.plot()

#Major_category vs part time
plt.figure(figsize=(10,10))
plt.title("Major_category vs Part_time")
sns.barplot(data=college_Majors,x="Part_time",y="Major_category")
plt.plot()

#change the wording here
#finding the relationships between variables that show the influence of one variable
on the other, especially on the target
#univariate regression
plt.figure(figsize=(100,100))
sns.pairplot(college_Majors, kind = 'reg', corner =True,
plot_kws={'line_kws':{'color':'cyan'}})
plt.show()

#testing our model in Unemployed vs Men
Y = college_Majors['Men']
X=college_Majors['Unemployed']
X.head()

X = sm.add_constant(X)
X.head(15)

model = sm.OLS(Y, X)
```

```
model_fitted = model.fit()
model_fitted.summary()


#fit our model to determin if Unemployed has a very strong relationship with men
sm.graphics.plot_fit(model_fitted,1, vlines=False);


#testing our model in Unemployed vs women
Y = college_Majors['Women']
X=college_Majors['Unemployed']
X.head()

X = sm.add_constant(X)
X.head(15)

model = sm.OLS(Y, X)
model_fitted = model.fit()
model_fitted.summary()

#fit our model to determin if Unemployed has a very strong relationship with women
sm.graphics.plot_fit(model_fitted,1, vlines=False);
```

References

1. Casselman, B. (2014, September 12). *The economic guide to picking a college major*. FiveThirtyEight. https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/
2. *American Association of engineering societies (AAES) awards*. (n.d.). AIME | The American Institute of Mining, Metallurgical, and Petroleum Engineers. https://aimehq.org/what-we-do/awards/american-association-engineering-societies-aaes-awards
3. Data/college-majors at master · fivethirtyeight/data. (n.d.). GitHub. https://github.com/fivethirtyeight/data/tree/master/college-majors