Final Project — Milestone 3

ALY 6010 Probability Theory and Statistics

RAHUL AVINASH JADHAV

Northeastern University

College of Professional Studies, Northeastern University, Boston, MA 02115

Contact: jadhav.ra@northeastern.edu

Submitted to Professor: Prof. Roy Wada

Date of submission: 12/19/2021

Introduction

The dataset we are using for our milestone is NLSY1979_1994. The dataset consists of 8847 observations and 37 variables. The dataset contains information about the individual like their age, gender, income, majors of degree they hold, marital status etc.

We are going to perform statistical analysis of the dataset. Choose one variable and see the relationship it has with income variable. Perform hypothesis test like one sample and 2 sample tests. And perform regression model.

The Hypothesis test will let us know the relationship between the groups.

The regression model helps us to define the relationship between the variables and to know if the variable has any effect on the independent variable.

Analysis

Table 1: Summary table of the Dataset used.

(Note the count is 8847)

	mean	standard_deviation	min	max	standard_error
GENDER*	1.5	0.5	1	2	0.01
Age	33.4	2.21	30	37	0.02
RACE*	2.2	0.88	1	3	0.01
HAVING_HEALTHPLAN*	2.6	0.8	1	3	0.01
REGION_*	2.8	1.36	1	5	0.01
URBAN_RURAL_*	2.6	0.79	1	3	0.01
MARSTAT_KEY_*	2.3	0.8	1	6	0.01
WKSUEMP_PCY_	2.6	7.74	0	52	0.08
EDU_DEGREE*	4	1.57	1	8	0.02
MAJOR_1_*	17.8	8.93	1	26	0.09
INCOME_	18677.7	18953.83	0	101653	201.51
NET_WORTH_	53514.8	137001.9	0	946749	1456.56

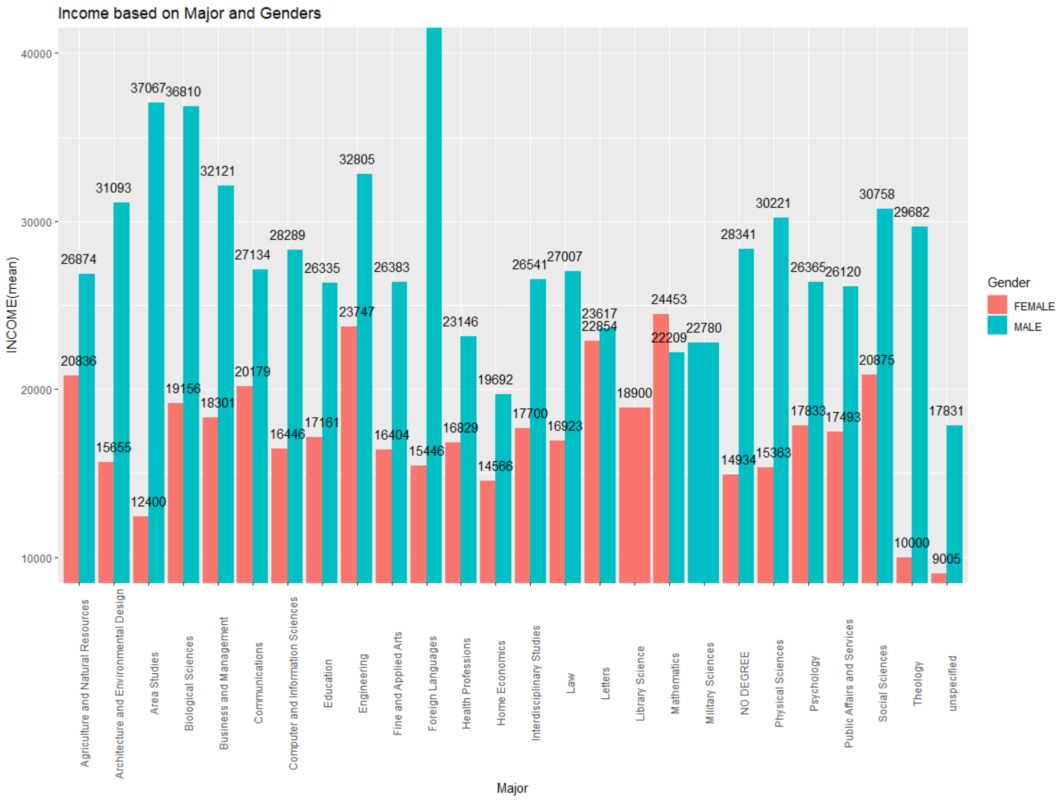
The above table provides statistically information about the Dataset. We can observe that the mean age group is 33. The mean income of the dataset is 18677.7 and mean net worth is 53514.8

Table 2: Statistical information about the dataset grouped by Majors and gender

	MAJOR_1_	GENDER	income	networth	count	proportion
1	unspecified	FEMALE	9005	30009	1923	0.462
2	unspecified	MALE	17831	31877	2242	0.538
3	Theology	FEMALE	10000	642999	3	0.214
4	Theology	MALE	29682	22441	11	0.786
5	Social Sciences	FEMALE	20875	140791	83	0.488
6	Social Sciences	MALE	30758	98459	87	0.512
7	Public Affairs and Services	FEMALE	17493	54607	74	0.481
8	Public Affairs and Services	MALE	26120	46011	80	0.519
9	Psychology	FEMALE	17833	67829	109	0.732
10	Psychology	MALE	26365	40113	40	0.268
11	Physical Sciences	FEMALE	15363	89212	22	0.301
12	Physical Sciences	MALE	30221	92846	51	0.699
13	NO DEGREE	FEMALE	14934	83577	108	0.482
14	NO DEGREE	MALE	28341	111856	116	0.518
15	Military Sciences	MALE	22780	44480	5	1
16	Mathematics	FEMALE	24453	43652	18	0.439
17	Mathematics	MALE	22209	24948	23	0.561
18	Library Science	FEMALE	18900	0	1	1
19	Letters	FEMALE	22854	26736	36	0.571
20	Letters	MALE	23617	32962	27	0.429
21	Law	FEMALE	16923	63646	29	0.492
22	Law	MALE	27007	66644	30	0.508
23	Interdisciplinary Studies	FEMALE	17700	68011	126	0.383
24	Interdisciplinary Studies	MALE	26541	61445	203	0.617
25	Home Economics	FEMALE	14566	82679	33	0.717
26	Home Economics	MALE	19692	32788	13	0.283
27	Health Professions	FEMALE	16829	57908	428	0.849
28	Health Professions	MALE	23146	63303	76	0.151

29	Foreign Languages	FEMALE	15446	119054	20	0.8
30	Foreign Languages	MALE	46726	48440	5	0.2
31	Fine and Applied Arts	FEMALE	16404	61032	85	0.47
32	Fine and Applied Arts	MALE	26383	66800	96	0.53
33	Engineering	FEMALE	23747	106192	35	0.101
34	Engineering	MALE	32805	79858	313	0.899
35	Education	FEMALE	17161	71148	273	0.705
36	Education	MALE	26335	76665	114	0.295
37	Computer and Information Sciences	FEMALE	16446	63447	192	0.542
38	Computer and Information Sciences	MALE	28289	55831	162	0.458
39	Communications	FEMALE	20179	77585	50	0.439
40	Communications	MALE	27134	48296	64	0.561
41	Business and Management	FEMALE	18301	65453	713	0.612
42	Business and Management	MALE	32121	93270	452	0.388
43	Biological Sciences	FEMALE	19156	132119	63	0.438
44	Biological Sciences	MALE	36810	117761	81	0.562
45	Area Studies	FEMALE	12400	22450	5	0.625
46	Area Studies	MALE	37067	163228	3	0.375
47	Architecture and Environmental Design	FEMALE	15655	107375	20	0.339
48	Architecture and Environmental Design	MALE	31093	50507	39	0.661
49	Agriculture and Natural Resources	FEMALE	20836	40468	14	0.215
50	Agriculture and Natural Resources	MALE	26874	80000	51	0.785

Bar Plot 1: Income based on Majors.



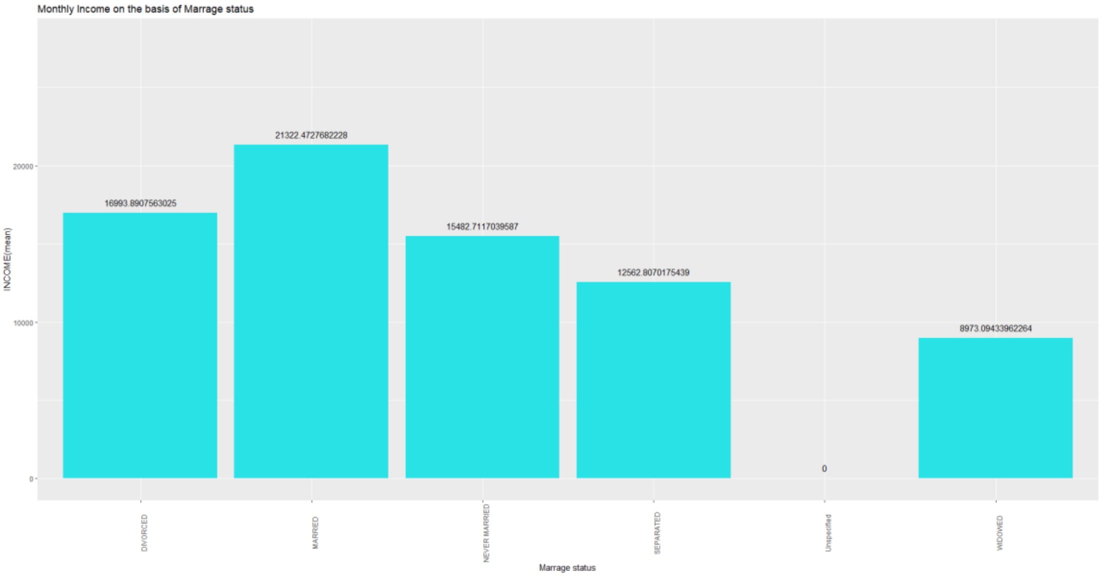
The above table and plot provide comparison between the income based on Majors and Gender. If we observe Properly most of the majors pay Male more than what they pay females.

- # INCOME BASED ON MARITAL STATUS

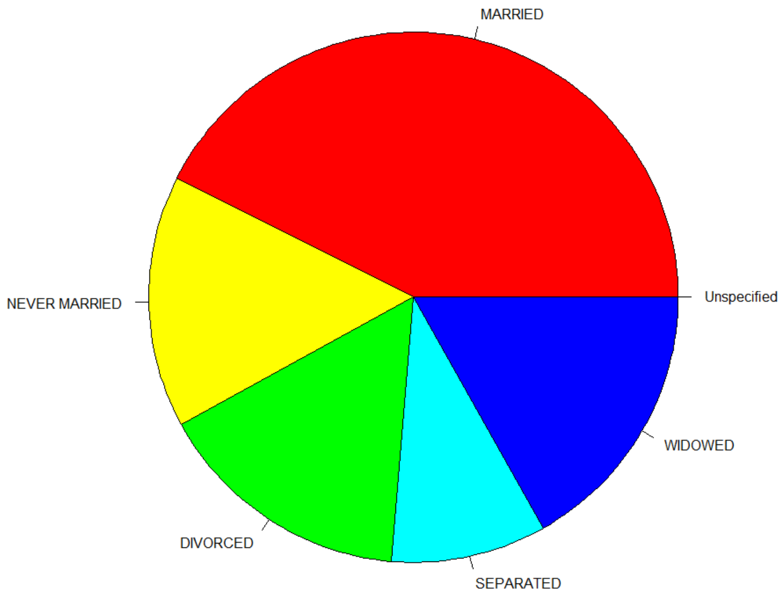
Table 3: Summary of income based on marital status

MARSTAT_KEY_	mean(INCOME_)	mean(NET_WORTH_)	count	proportion
MARRIED	21322.47277	75724.85708	4884	0.552
NEVER MARRIED	15482.7117	27312.46816	2324	0.263
DIVORCED	16993.89076	27887.20635	1071	0.121
SEPARATED	12562.80702	16916.64912	513	0.058
WIDOWED	8973.09434	29910.62264	53	0.006
Unspecified	0	0	2	0

Bar plot 2: Monthly income based on marital status



Pie 1: Net worth based on marital status



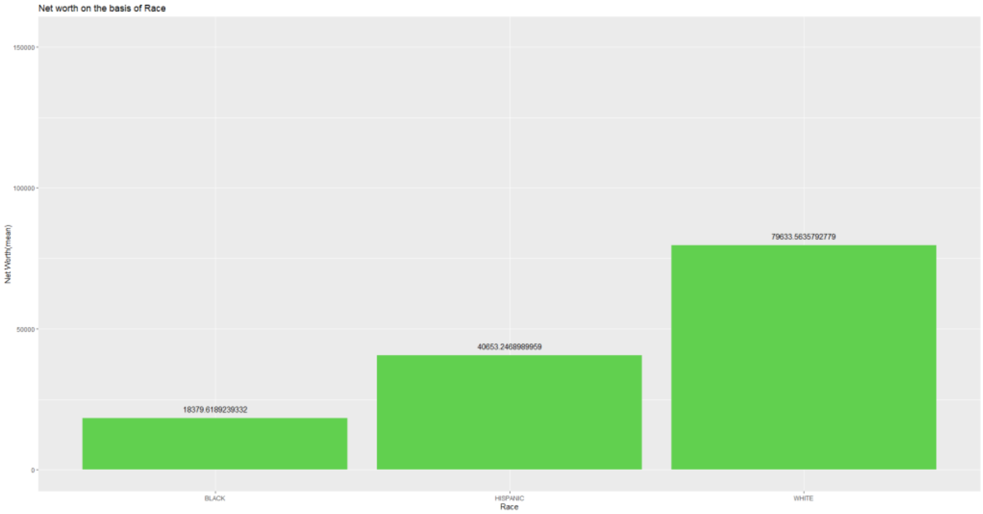
From table one we can see that the proportion of income data is more of married person. If closely observed we can also say that married couple have more net worth and income than rest marital status whereas the income and net worth of Separated people are the least.

- ## Income based on race

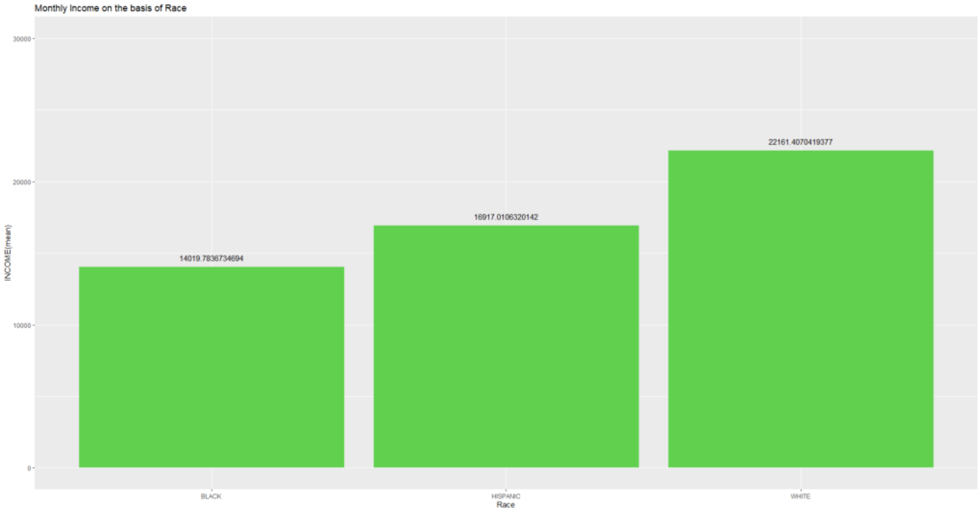
Table 4: Summary of income based on Race

	RACE	mean(INCOME_)	mean(NET_WORTH_)	count	proportion
1	WHITE	22161.41	79633.56	4459	0.504
2	BLACK	14019.78	18379.62	2695	0.305
3	HISPANIC	16917.01	40653.25	1693	0.191

Bar plot 3: Net worth based on Race



Bar plot 4: Income based on Race



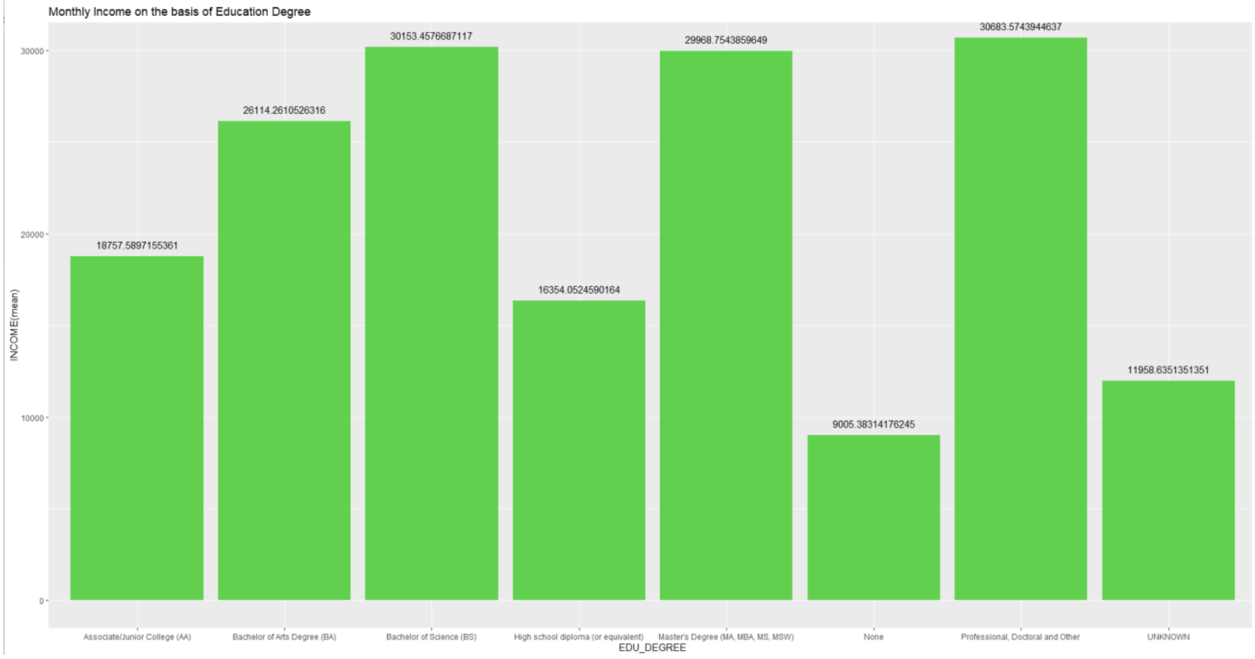
From the above table and bar plot we can say that people of race white have the highest income and net worth whereas people of race black have the least.

- INCOME BASED ON EDUCATION QUALIFICATION

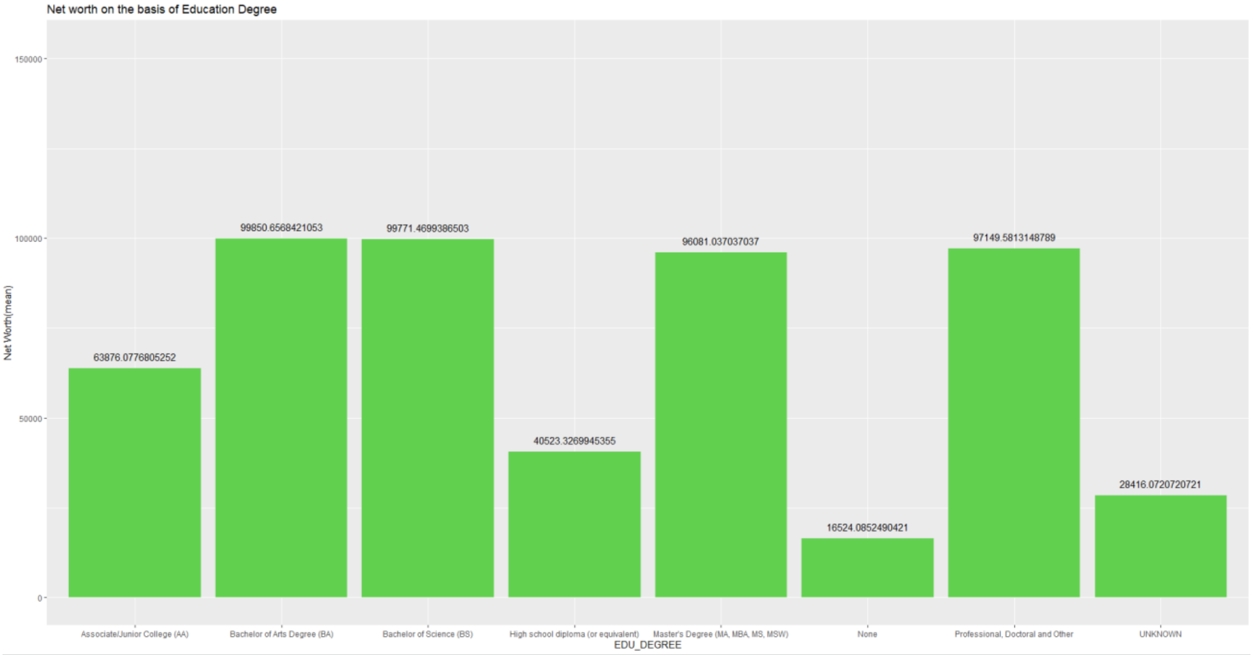
Table 5: Summary of income based on educational qualification

	EDU_DEGREE	mean(INCOME_)	mean(NET_WORTH_)	count	proportion
1	High school diploma (or equivalent)	16354.05	40523.33	4575	0.517
2	None	9005.383	16524.09	1044	0.118
3	Associate/Junior College (AA)	18757.59	63876.08	914	0.103
4	Bachelor of Science (BS)	30153.46	99771.47	815	0.092
5	Master's Degree (MA, MBA, MS, MSW)	29968.75	96081.04	513	0.058
6	Bachelor of Arts Degree (BA)	26114.26	99850.66	475	0.054
7	Professional, Doctoral and Other	30683.57	97149.58	289	0.033
8	UNKNOWN	11958.64	28416.07	222	0.025

Bar plot 5: Income based on Educational Qualification



Bar plot 6: Net worth based on Educational Qualification



From the above plots of income based on educational qualification we can see that individual with professional Doctoral or other degree have the highest income where as individual with no degree have the least income. But when we check the net worth plot, we can see that individual with bachelor of arts degree have the highest net worth whereas the individual with no degree have the least income.

Table 6: One sample t test with alternative as two sided on income based on Majors and gender.

	Gender	count	Mean income	statistic	p.value	conf.low	conf.high
1	ALL	49	22788.12	21.76	0	20683.36	24892.88
2	FEMALE	24	17338.36	23.18	0	15794.82	18881.9
3	MALE	24	28237.88	23.49	0	25756.91	30718.85

H0 = Mean income of each group is equal to 0.

H1 = Mean income of each group is unequal to 0.

The t-test of the mean were conducted against the null hypothesis that the sample mean was equal to 0. Table 3 displays the results of the t test for the income of Majors as whole and by group. The female income and male income mean was statistically different from 0 at p as 0. Thus, we will reject the null hypothesis

Table 7: Welch Two Sample t-test on mean income based on Gender with alternative and two sided

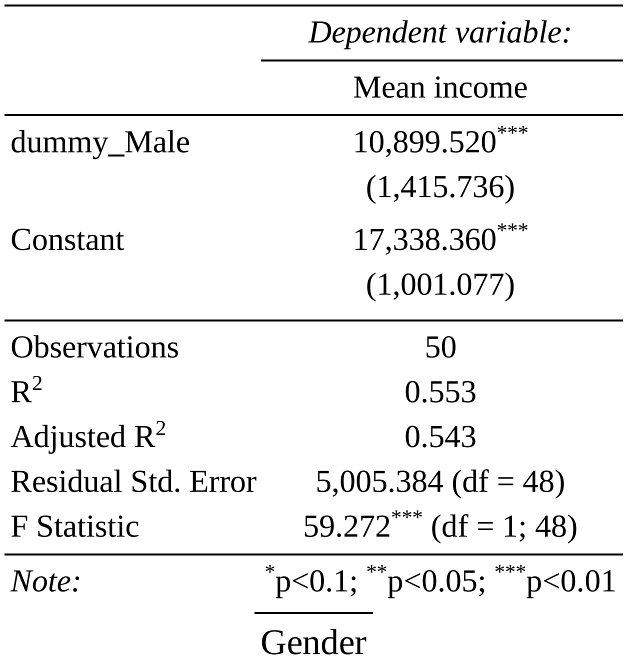
estimate	Mean income(Female)	Mean income(Male)	count	statistic	p.value	conf.low	conf.high
-10899.5	17338.36	28237.88	40.15854	-7.7	0	-13760.5	-8038.56

H0 = Mean income of the groups is equal to 0.

H1 = Mean income of the groups is unequal to 0.

The two-sample t-test of the mean were conducted against the null hypothesis that the sample mean was equal to 0. Table 4 displays the results of the t test for the income of Genders. The female income and male income mean was statistically different from 0 at p as 0. Thus, we will reject the null hypothesis

Table 8: Regression on income using dummy variable Gender



	Dependent variable:
	Mean income
dummy_Male	10,899.520***
	(1,415.736)
Constant	17,338.360***
	(1,001.077)
Observations	50
R^2	0.553
Adjusted R^2	0.543
Residual Std. Error	5,005.384 (df = 48)
F Statistic	59.272*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

Gender

From the above table we can see that income of male is 10,899.520 more than income of female for every increment. The r squared value is 0.55. our regression model is 55% fit for our observation.

Table 9: Regression on income based on Gender and major

	Dependent variable:
	Mean income
GENDERMALE	11,192.000***
	(1,458.876)
MAJOR_1_Architecture and Environmental Design	-481.000
	(5,053.695)
MAJOR_1_Area Studies	878.500
	(5,053.695)
MAJOR_1_Biological Sciences	4,128.000
	(5,053.695)
MAJOR_1_Business and Management	1,356.000
	(5,053.695)
MAJOR_1_Communications	-198.500
	(5,053.695)
MAJOR_1_Computer and Information Sciences	-1,487.500
	(5,053.695)
MAJOR_1_Education	-2,107.000
	(5,053.695)
MAJOR_1_Engineering	4,421.000
	(5,053.695)
MAJOR_1_Fine and Applied Arts	-2,461.500
	(5,053.695)
MAJOR_1_Foreign Languages	7,231.000
	(5,053.695)
MAJOR_1_Health Professions	-3,867.500
	(5,053.695)
MAJOR_1_Home Economics	-6,726.000
	(5,053.695)
MAJOR_1_Interdisciplinary Studies	-1,734.500
	(5,053.695)
MAJOR_1_Law	-1,890.000
	(5,053.695)
MAJOR_1_Letters	-619.500

	(5,053.695)
MAJOR_1_Library Science	641.000
	(6,232.321)
MAJOR_1_Mathematics	-524.000
	(5,053.695)
MAJOR_1_Military Sciences	-6,671.000
	(6,232.321)
MAJOR_1_NO DEGREE	-2,217.500
	(5,053.695)
MAJOR_1_Physical Sciences	-1,063.000
	(5,053.695)
MAJOR_1_Psychology	-1,756.000
	(5,053.695)
MAJOR_1_Public Affairs and Services	-2,048.500
	(5,053.695)
MAJOR_1_Social Sciences	1,961.500
	(5,053.695)
MAJOR_1_Theology	-4,014.000
	(5,053.695)
MAJOR_1_unspecified	-10,437.000*
	(5,053.695)
Constant	18,259.000***
	(3,647.190)
Observations	50
R^{2}	0.781
Adjusted R^{2}	0.534
Residual Std. Error	5,053.695 (df = 23)
F Statistic	3.163*** (df = 26; 23)
Note:	*p<0.1; **p<0.05; ***p<0.01

MAJOR GENDER

From the above regression table, we can see that income of Male is 11,192 different from that of income of Female. All the major's income are compared with Agriculture and natural resources major. We can see that Architecture and Environmental Design income is -481 less for every increment. The r squared value is 0.78. the regression model is 78% fit for our observations.

Conclusion

- We learned to perform t test and 2 sample t test on the dataset. T test can help us to know if the mean of the sample is similar to the population.
- We performed Regression testing on the dataset.
- We can see that there is positive relationship between income and Majors that higher the Majors level more the income.
- We also learnt the importance of dummy variables.

Bibliography

- Kabacoff, R. (2011). *R in action: Data analysis and graphics with R*. Manning.
- https://statisticsglobe.com/replace-negative-values-by-zero-in-r
- https://stackoverflow.com/questions/46526833/calculate-mean-standarddeviation-n-etc-across-columns-and-create-new-data-f/46528063