UTM

UNIVERSITI TEKNOLOGI MALAYSIA

Probability and Statistics Analysis
(SECI2143)

Group 4 (STILL LIFE)
Section 02

Lecturer's name	Dr Nor Azizah
Date of submission	27/6/2022
Group	Group 4 (STILL LIFE)
Section	Section 02
Member 1	Oon Yee Sem (A21EC0122)
Member 2	Mohamad Azri Hadif Bin Mohammad Rizal (A21EC0054)
Member 3	Nur Shuhada Safiah binti Ayob (A21EC0114)
Member 4	Nurunnajwa binti Zulkifli (A21EC0121)

Table of Content

Introduction and Background

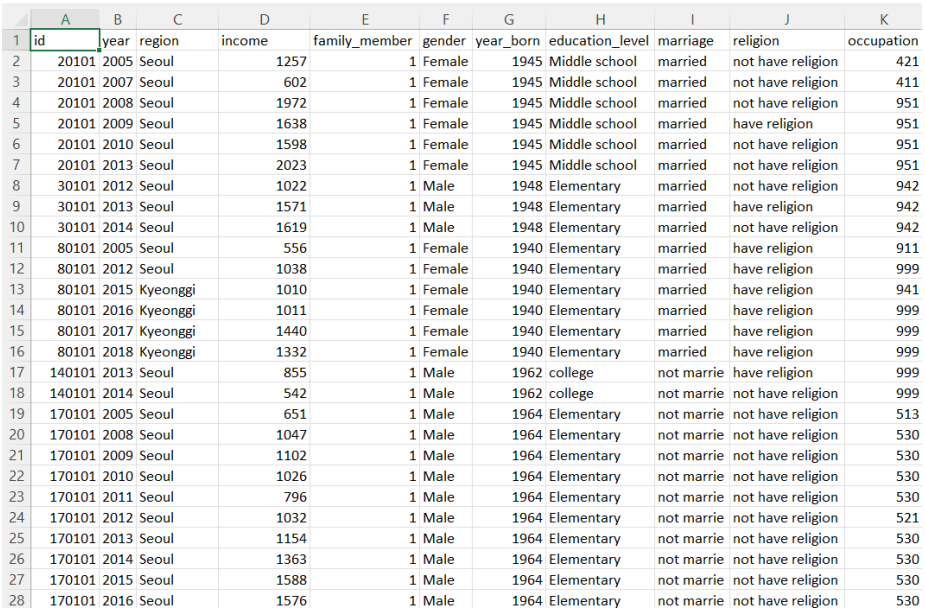
Money makes the world go around. Economies rely on the exchange of money for products and services. Economists define money, where it comes from and what it's worth. In order to survive, everyone must have money and in order to have money, we need to work; have a job. South Korea's economy has been on a turbulent ride in recent years with exports sagging and key industries suffering however South Korea actually rose on the world GDP rankings in 2015 when compared with other economies making it the eleventh largest economy in the world.

Hence, regarding the rise of South Korea's GDP, we have decided to observe what makes them achieve and still going strong with their economy as well as learn one or two things from this project on how to improve our country; Malaysia to be on par with fellow Asian countries. Therefore, the primary objective of this study is to analyze the relationship between income, gender, education, and other variables, as well as to demonstrate the existence of a relationship between the data using statistical analysis techniques. As a result, in order to achieve this objective, a handful of potential variables are carefully chosen, and a series of test analyses are conducted.

The dataset we have chosen regarding South Korea's economy is the Korean Income and Welfare is a secondary data source which facts and figures already recorded prior to the project retrieved from the Kaggle website. This data was collected by the Korea Welfare Panel Study (KOWEPS) where this organization was designed to provide a probability sample of South Korea's population. This dataset was collected to show 996 Koreans in terms of their id, region, family member, gender, year born, education level, marriage, religion and occupation. Data types that are obtained here are both qualitative and quantitative data. Moreover, levels of data measurement that are found are ratio scale, interval scale, ordinal scale and nominal scale.

Dataset

In Project 2 of Probability and Statistical Data Analysis, we use one data set that contains all the information necessary to conduct all required tests. This project requires us to conduct three tests: the hypothesis 1 or 2 sample test, the correlation test, and the regression test. Optional additional tests include the goodness-of-fit test, the Chi-square test of independence, and the ANOVA test. Both data sets that we used were found on Kaggle (https://www.kaggle.com/). Our data is about Korean Income and Welfare.



	A	B	C	D	E	F	G	H	I	J	K
1	id	year	region	income	family_member	gender	year_born	education_level	marriage	religion	occupation
2	20101	2005	Seoul	1257	1	Female	1945	Middle school	married	not have religion	421
3	20101	2007	Seoul	602	1	Female	1945	Middle school	married	not have religion	411
4	20101	2008	Seoul	1972	1	Female	1945	Middle school	married	not have religion	951
5	20101	2009	Seoul	1638	1	Female	1945	Middle school	married	have religion	951
6	20101	2010	Seoul	1598	1	Female	1945	Middle school	married	not have religion	951
7	20101	2013	Seoul	2023	1	Female	1945	Middle school	married	not have religion	951
8	30101	2012	Seoul	1022	1	Male	1948	Elementary	married	not have religion	942
9	30101	2013	Seoul	1571	1	Male	1948	Elementary	married	have religion	942
10	30101	2014	Seoul	1619	1	Male	1948	Elementary	married	not have religion	942
11	80101	2005	Seoul	556	1	Female	1940	Elementary	married	have religion	911
12	80101	2012	Seoul	1038	1	Female	1940	Elementary	married	have religion	999
13	80101	2015	Kyeonggi	1010	1	Female	1940	Elementary	married	have religion	941
14	80101	2016	Kyeonggi	1011	1	Female	1940	Elementary	married	have religion	999
15	80101	2017	Kyeonggi	1440	1	Female	1940	Elementary	married	have religion	999
16	80101	2018	Kyeonggi	1332	1	Female	1940	Elementary	married	have religion	999
17	140101	2013	Seoul	855	1	Male	1962	college	not marrie	have religion	999
18	140101	2014	Seoul	542	1	Male	1962	college	not marrie	not have religion	999
19	170101	2005	Seoul	651	1	Male	1964	Elementary	not marrie	not have religion	513
20	170101	2008	Seoul	1047	1	Male	1964	Elementary	not marrie	not have religion	530
21	170101	2009	Seoul	1102	1	Male	1964	Elementary	not marrie	not have religion	530
22	170101	2010	Seoul	1026	1	Male	1964	Elementary	not marrie	not have religion	530
23	170101	2011	Seoul	796	1	Male	1964	Elementary	not marrie	not have religion	530
24	170101	2012	Seoul	1032	1	Male	1964	Elementary	not marrie	not have religion	521
25	170101	2013	Seoul	1154	1	Male	1964	Elementary	not marrie	not have religion	530
26	170101	2014	Seoul	1363	1	Male	1964	Elementary	not marrie	not have religion	530
27	170101	2015	Seoul	1588	1	Male	1964	Elementary	not marrie	not have religion	530
28	170101	2016	Seoul	1576	1	Male	1964	Elementary	not marrie	not have religion	530

Figure 1.0 : Korean Income and Welfare

For Korean Income data set, we have eleven variables: ID, year, region, income, family member, gender, year born, education level, marriage, religion and occupation. Chosen variables to run the tests are income, gender, family member, year born, education level and marriage.

Data Description

Variable	Type of Variable	Measurement Level
Income	Quantitative	Ratio
Gender	Qualitative	Nominal
Family Member	Quantitative	Ratio
Year Born	Qualitative	Interval
Education Level	Qualitative	Ordinal
Marriage	Qualitative	Nominal

Table 1.0 : Data Description

Data Analysis

1. Hypothesis Two Sample Test

For this test, we use two variables: gender and income. We will test whether the mean of income of male is different than the mean of income of female at 95% confidence level, assuming unequal variances. Frequency(n), mean($\bar{x}$) and standard deviation(s) from our data are calculated and recorded.

$\bar{x}_1 = 3028.117$	$\bar{x}_2 = 2976.071$
$s_1 = 2761.886$	$s_2 = 1926.428$
$n_1 = 652$	$n_2 = 344$

1. Hypothesis statement:

   $H_0 : \mu_1 = \mu_2$
   $H_1 : \mu_1 \neq \mu_2$
   Where $\mu_1$equals the mean of income for female, and μ_2 equals the mean of income for male.

2. Given 95% confidence level, $\alpha = 0.05$. The test statistics, t_0 can be calculated by formula below:

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}}$$

Figure 2.0 : t_0 formula

By using RStudio, test statistics, $t_0 = 0.347$

3. Calculate the degree of freedom, v by formula below:

$$v = \frac{\left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)^2}{\frac{\left(S_1^2/n1\right)^2}{n1-1} + \frac{\left(S_2^2/n2\right)^2}{n2-1}}.$$

Figure 3.0 : Degree of Freedom, v formula

By using RStudio, degree of freedom, v = 920.16
920.16 ≈ 920; v = 920

4. Therefore, using α = 0.05, we reject H_0 if $t_0 > t_{0.025,\,920}$ and if $t_0 < t_{-0.025,\,920}$. Critical value, $t_{0.025,\,920} = 1.647$ and $t_{-0.025,\,920} = -1.647$ while p-value = 0.7284

5. **Description:** Based on the analysis above, we found that there is no difference between mean of female income and mean of male income. This concludes that income between males and females in South Korea don't have a huge difference nowadays. One of the factors is that recently women in South Korea also take part in joining the workforce in their country. In addition, South Korea has enforced their rule as gender equality is included as one of the six basic pillars for development in South Korea's 'Framework Act on International Development Cooperation'. KOICA's Mid-term Sectoral Strategy 2021-2025 focuses on three strategic objectives for gender: economic empowerment; social status; and basic rights.

6. **Analysis:** Since test statistic $t_0 = 0.347$, $t_0 > t_{-0.025,\,920}$ and $t_0 < t_{0.025,\,920}$, we failed to reject H_0 at α = 0.05.

7. **Conclusion:** There is no sufficient evidence that the mean income of male is different from the mean income of female

2. Correlation Test

To observe the strength of relationship between two variables; income and family member
data used are filter for income narrow it to 100<y<24000

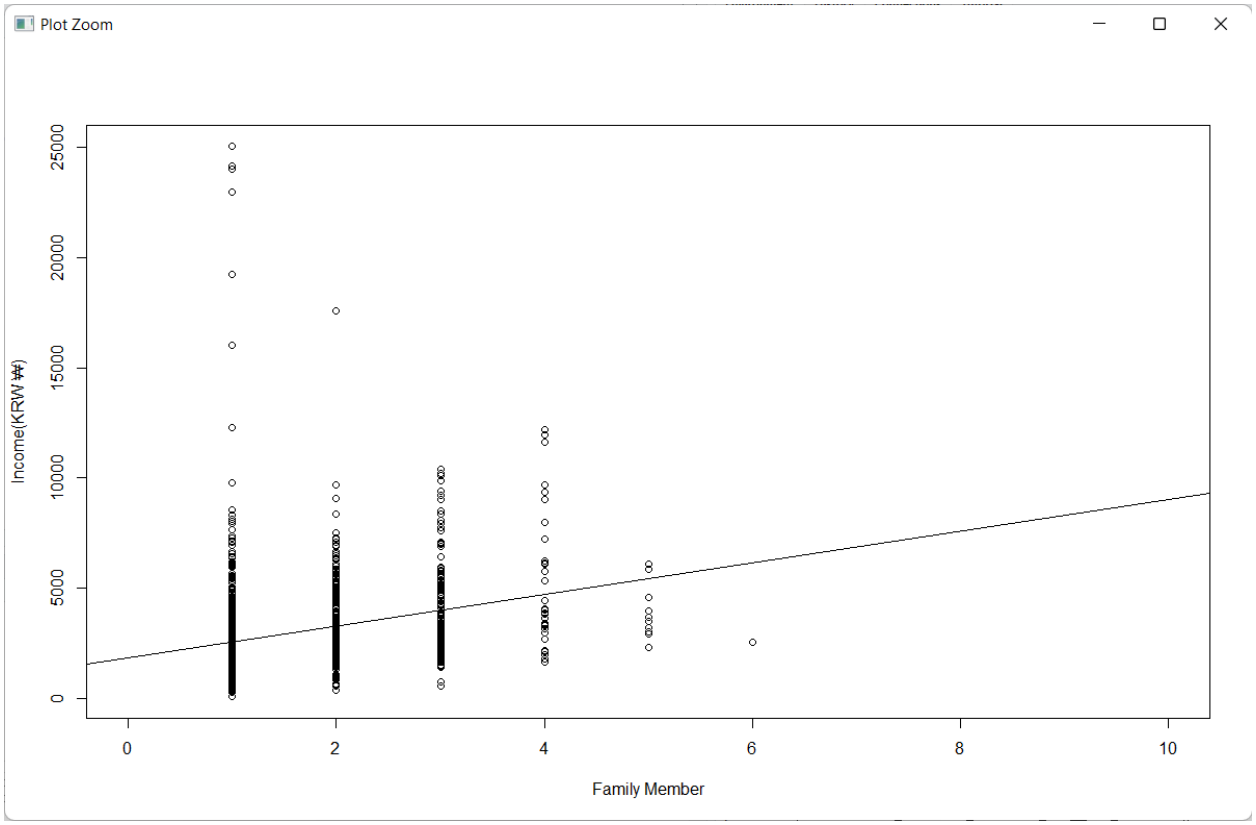


Figure 3.0 : Scatter plot Income vs Family Member

x - axis: Family Member

y - axis: Income(KRW ₩)

H_0: p = 0

H_1: p ≠ 0

Pearson's product-moment correlation

$$r = ([\sum xy - (\sum x \sum y)]/n) / \sqrt{[(\sum x^2) - (\sum x)^2/n]\,[(\sum y)^2 - (\sum y)^2/n]}$$

r = 0.2656558 (weak positive linear relationship)

Critical value

95%: $\alpha = 0.05$

Degree of freedom = 994 - 2 = 992

$t_{(0.025,992)} = 1.960$

Test statistics: $t = r/\left(\sqrt{(1 - r^2)/(n - 2)}\right)$
$= 8.679$

Description: Based on the analysis above, we found that the strength between Income and Family Member is weak positive correlation. It can be seen that the income increases as the family member increases. A scatter plot and correlation analysis of the data indicates that there is a positive relationship between the number of income and family members. Since the correlation is closer to 0, it has a weaker positive linear relationship.

Analysis: Since $t = 8.679 > t_{(0.025,992)} = 1.960$, we reject H_0.

Conclusion: There is sufficient evidence to conclude that there is a linear relationship between income and family members at the 5% level of significance / 95% confidence interval.

3. Regression Test

To predict the value of income based on the year born

Dependent value : income Independent value : year born

$H_0 : \beta 1 = 0$

$H_1 : \beta 1 \neq 0$

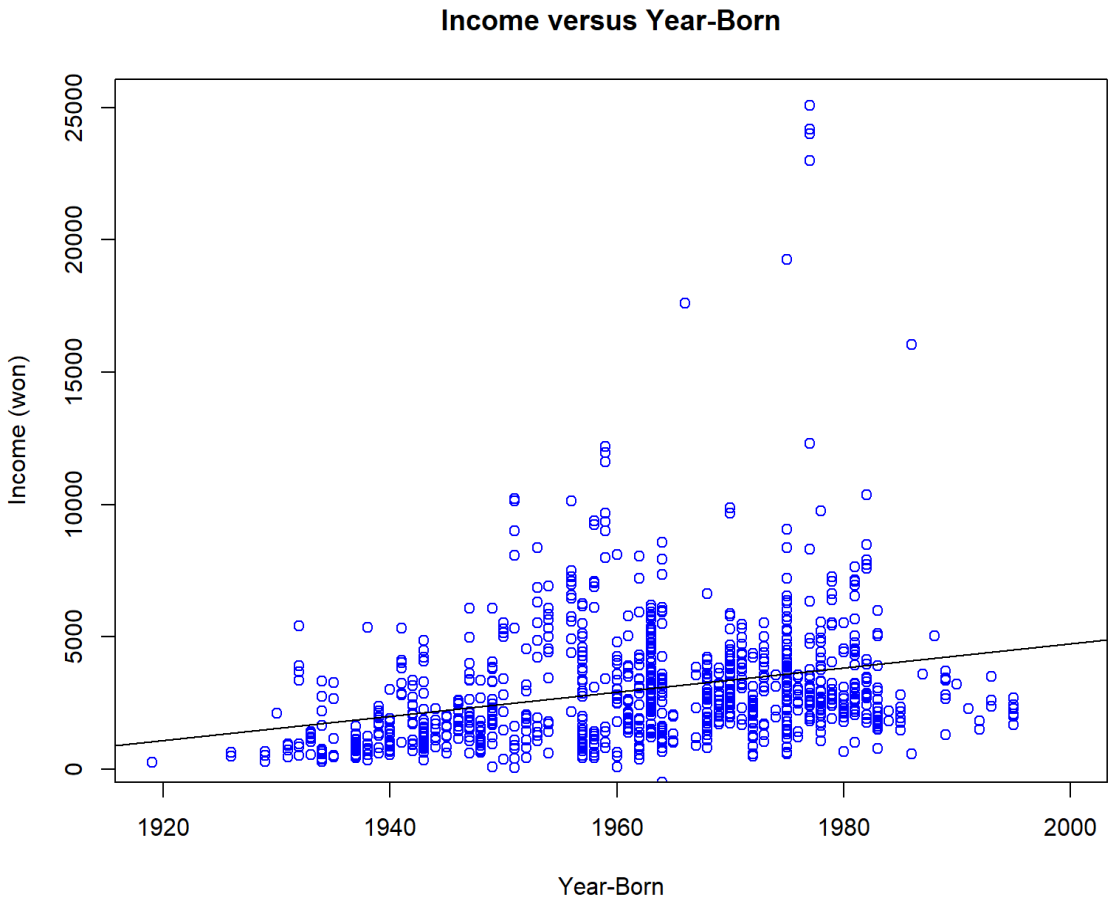


Figure 4.0 : Regression of Income versus Year-Born

Let x - year-born (independent value)
Let y - income (dependent value)

Calculate Least Square Equation

$$b_1 = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

$= -86674.14$ $= 45.73$

$$\hat{y} = b_1 x + b_0$$

$= 45.73x - 86674.14$

Calculate Coefficient of Determination

$$R^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

$= 0.078$

Calculate Standard Error of Estimate

$$S_\varepsilon = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - k - 1}}$$

$= 2406$

Calculate the Standard Deviation of Regression Slope

$$S_{b1} = \frac{S_\varepsilon}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}$$

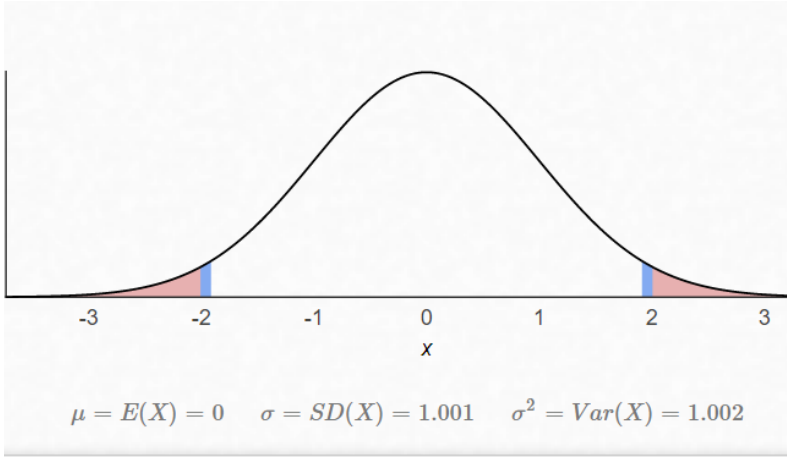
$= 4.987$

Calculate the Test Statistic

$$t = \frac{b_1 - \beta_1}{S_{b1}}$$

$= 9.170$

P-value = 0



(Critical Region)

$$\alpha = 0.05$$

$$df = n - 2$$
$$= 994$$

$$c.v.t_{df,-\alpha/2} = -1.96 \qquad c.v.t_{df,\alpha/2} = 1.96$$

Description - Based on the analysis above, we observed that the people with the highest income are approximately born between 1960 to 1990. It is because these years (1960 to 1990) are the golden ages of Korea in which Korea was declared as one of the Four Asian Dragons with rapid industrialization and maintained exceptionally high growth rates of more than 7 percent per year.Meanwhile, we also found that the people with the low income are approximately born between 1990 to 1930 because second world war and high inflation rate occured at that time and most of the people lost their jobs and lead to low income.

Analysis - Since 1.96 < statistical value 9.170 and P-value < 0.05. Reject H_0 at a significant level of 0.05

Conclusion - There is sufficient enough to conclude that the variable year-born affects the income of korean

4. Goodness-of-Fit Test

To test whether the educational level with the same proportion

Let :No education = n, Elementary = e, Middle school =m
High school = h, College = c , Degree = d
Master = m , Doctoral Degree = doc

Claim : Educational levels are in the same proportion.Therefore,

$$P_n = P_e = P_m = P_h = P_c = P_d = P_m = P_{doc}$$

$H_0 : P_n = P_e = P_m = P_h = P_c = P_d = P_m = P_{doc}$

H_1 : At least one of the eight proportions is different from others

$$E = \frac{n}{k} = \frac{996}{8} = 124.5$$

Educational levels	Observed Count(O)	Expected Count(E)	[O - E)]2/ E
No education	85	124.5	12.53
Elementary	160	124.5	10.12
Middle school	126	124.5	0.018
High school	293	124.5	228.05
College	114	124.5	0.8855
Degree	178	124.5	22.99
Master	27	124.5	76.35
Doctoral Degree	13	124.5	99.86
			$\Sigma X^2 = 450.81$

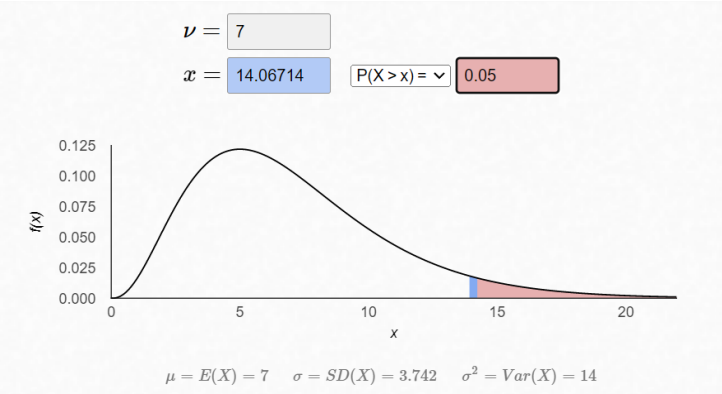
Table 2.0 : Observed Counts, Expected Count, Test Statistical Value

$$X^2 = \Sigma \frac{(O-E)^2}{E}$$
$$= \quad 450.81$$

$$\alpha = 0.05$$

$$df = n - 1$$
$$= 7$$

P-value = 2.2×10^{-16}



(Critical Region)

$$X^2_{df,\alpha} = X^2_{7,0.05}$$
$$= 14.06$$

Description - Based on the analysis above,we found that the proportions of educational level of Koreans are different.We found that the people with highest educational level, Doctoral Degree are the lowest with only have 13 people and the people with low educational level,High school level have the highest number of people with 295 people.

Analysis - Since critical value 14.06 < test statistical value 450.81 and P-value < 0.05. Reject H_0 at a significant level of 0.05

Conclusion - We reject the claim that the educational level with equal proportions (frequency) on the 8 different educational levels

5. Chi Square Independent Test

To test whether the educational level and marriage is independent or not

H_0 : Educational level is independent of marriage

H_1 : Educational level is not independent of marriage

	Marriage		
Educational Level	Married	Not Married	Total
No education	85	0	85
Elementary	130	30	160
Middle school	91	35	126
High school	115	178	293
College	12	102	114
Degree	6	172	178
Master	0	27	27
Doctoral Degree	0	13	13
Total	439	557	996

Table 3.0 : Education Level and Marriage

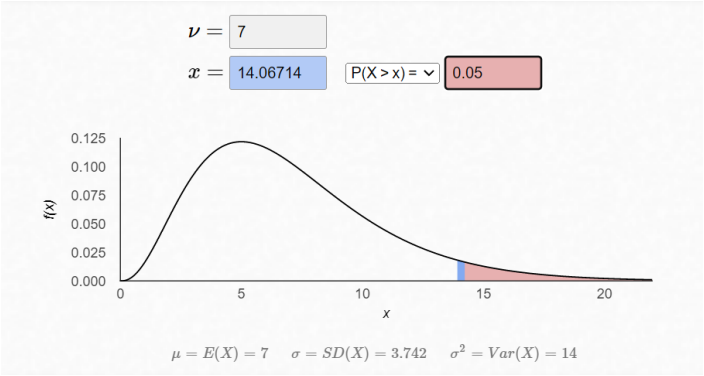
Cell	Observed Count(O_i)	Expected Count(E_i)	[O_i - E_i)]2/ E_i
1,1	85	37.465	60.312
1,2	0	47.535	47.535
2,1	130	70.522	50.163
2,2	30	89.478	39.536
3,1	91	55.536	22.646
3,2	35	70.463	17.849
4,1	117	129.144	1.549
4,2	178	163.856	1.220
5,1	12	50.297	29.113
5,2	104	63.753	22.945
6,1	6	78.456	66.914
6,2	172	99.544	52.739
7,1	0	11.900	11.901
7,2	27	15.099	9.379
8,1	0	5.730	5.730
8,2	13	7.270	4.516
			$\Sigma X^2 = 444.05$

Table 4.0 : Observed Counts, Expected Count, Test Statistical Value

$\alpha = 0.05$

$df = (r - 1)(c - 1)$
$= 7$

$X^2 = \Sigma \frac{(O-E)^2}{E}$
$= 444.05$



(Critical Region)

P-value = 5.133×10^{-92}

$X^2_{df,\alpha} = X^2_{7,0.05}$
$= 14.06$

Description - Based on the analysis above,we found that the variable educational level of Korean affects the variable marriage status of Korean. For those who have a higher educational level (educational level higher or equal College level) most of them are not married. It is because most of them are financially independent and do not need to rely on their spouses.Besides that,people with higher educational level have higher requirements on their spouses compared to those who have lower educational level(educational level lower or equal to high school level).Moreover,people with higher educational level are more focusing on their studies, occupation or personal achievements so they have limited time to find spouses and date.

Analysis - Since critical value 14.06 < test statistical value 444.05 and P-value < 0.05. Reject H_0 at a significant level of 0.05

Conclusion - There is sufficient enough to conclude that the variable,Educational level is not independent to the variable marriage

Conclusion

In a nutshell, we have learned about the essential skill of performing data analysis based on a real dataset. We have chosen the Korean Income and Welfare dataset from Kaggle as it contains 11 variables and has 996 datasets after we clean it properly. We divided our tasks, and the result of the analysis has opened our eyes to the rapid development of the country South Korea from 2005 to 2018.

For data pre-processing, we implement data cleaning and data reduction. We found out that the data has five negative values of income. Also, we perform a boxplot to detect any extreme outliers. We deleted the negative value and outliers to enhance the efficiency of the data. Likewise, we reduced the column into 996 rows and only used six variables (income, gender, family member, year born, educational level, marriage.)

For two-sample hypothesis tests, we fail to reject the null hypothesis. Thus, the mean of male income is equal to the standard of female income. Even though South Korea has faced gender inequality since a long time ago, the government is trying to close the gender gap, from government support for paternity leave to the role of the private sector in boosting women's careers. For correlation, the number of family members has a weaker positive relationship with income. In addition, we found that as each year is born, the income also increases by regression. This especially happens to people in their 30's to 50's. While, for the goodness of fit tests, we know that the proportion of doctoral degrees is the lowest. Besides, the Chi-Square Independent Test depicts that education and marriage are not independent.

The most exciting finding from our results is the result of two sample hypothesis testing indicate that 6 basic pillars of South Korea economy in term on gender equality where male and femake can hold same position without discrimination. In addition, income is not dependent on family members. Besides, the person that was born between 1960 until 1990 has made a strong base towards a well developed South Korea as they have become one of Four Asian Dragons with rapid industrialization and maintained exceptionally high growth rates.

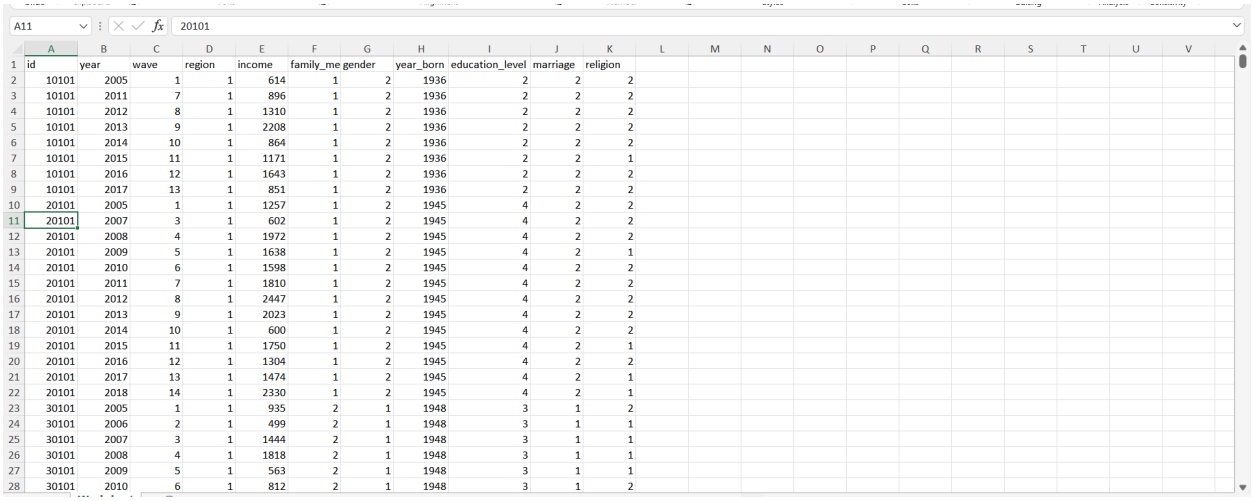
To conclude, we believe that every soft skill and hard skill that we gain from this project can help us increase our knowledge of data analysis and statistics as well as prepare ourselves for the world of work in the future. Thank you to Dr Azizah for helping and guiding us throughout this project.

Video Link

https://drive.google.com/file/d/1dhTIwPQsbDPxmQ5IDgvJ-dByuGosSdqJ/view?usp=sharing

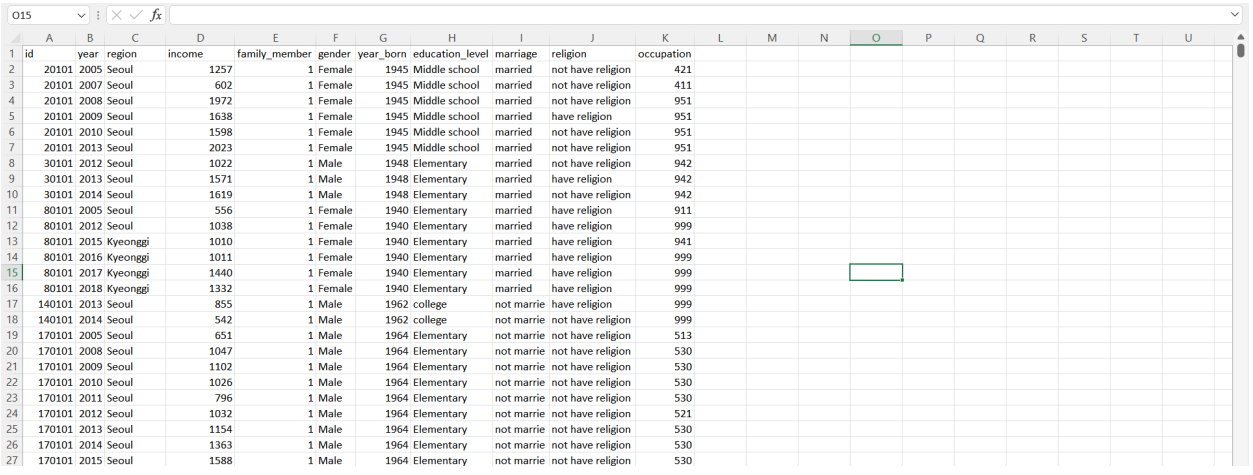
Appendix

1. Raw Data



id	year	wave	region	income	family_me	gender	year_born	education_level	marriage	religion
10101	2005	1	1	614	1	2	1936	2	2	2
10101	2011	7	1	896	1	2	1936	2	2	2
10101	2012	8	1	1310	1	2	1936	2	2	2
10101	2013	9	1	2208	1	2	1936	2	2	2
10101	2014	10	1	864	1	2	1936	2	2	2
10101	2015	11	1	1171	1	2	1936	2	2	1
10101	2016	12	1	1643	1	2	1936	2	2	2
10101	2017	13	1	851	1	2	1936	2	2	2
20101	2005	1	1	1257	1	2	1945	4	2	2
20101	2007	3	1	602	1	2	1945	4	2	2
20101	2008	4	1	1972	1	2	1945	4	2	2
20101	2009	5	1	1638	1	2	1945	4	2	1
20101	2010	6	1	1598	1	2	1945	4	2	2
20101	2011	7	1	1810	1	2	1945	4	2	2
20101	2012	8	1	2447	1	2	1945	4	2	2
20101	2013	9	1	2023	1	2	1945	4	2	2
20101	2014	10	1	600	1	2	1945	4	2	2
20101	2015	11	1	1750	1	2	1945	4	2	1
20101	2016	12	1	1304	1	2	1945	4	2	2
20101	2017	13	1	1474	1	2	1945	4	2	1
20101	2018	14	1	2330	1	2	1945	4	2	1
30101	2005	1	1	935	2	1	1948	3	1	2
30101	2006	2	1	499	2	1	1948	3	1	1
30101	2007	3	1	1444	2	1	1948	3	1	1
30101	2008	4	1	1818	2	1	1948	3	1	1
30101	2009	5	1	563	2	1	1948	3	1	1
30101	2010	6	1	812	2	1	1948	3	1	2

2. Cleaned Data



id	year	region	income	family_member	gender	year_born	education_level	marriage	religion	occupation
20101	2005	Seoul	1257	1	Female	1945	Middle school	married	not have religion	421
20101	2007	Seoul	602	1	Female	1945	Middle school	married	not have religion	411
20101	2008	Seoul	1972	1	Female	1945	Middle school	married	not have religion	951
20101	2009	Seoul	1638	1	Female	1945	Middle school	married	have religion	951
20101	2010	Seoul	1598	1	Female	1945	Middle school	married	not have religion	951
20101	2013	Seoul	2023	1	Female	1945	Middle school	married	not have religion	951
30101	2012	Seoul	1022	1	Male	1948	Elementary	married	not have religion	942
30101	2013	Seoul	1571	1	Male	1948	Elementary	married	have religion	942
30101	2014	Seoul	1619	1	Male	1948	Elementary	married	not have religion	942
80101	2005	Seoul	556	1	Female	1940	Elementary	married	have religion	911
80101	2012	Seoul	1038	1	Female	1940	Elementary	married	have religion	999
80101	2015	Kyeonggi	1010	1	Female	1940	Elementary	married	have religion	941
80101	2016	Kyeonggi	1011	1	Female	1940	Elementary	married	have religion	999
80101	2017	Kyeonggi	1440	1	Female	1940	Elementary	married	have religion	999
80101	2018	Kyeonggi	1332	1	Female	1940	Elementary	married	have religion	999
140101	2013	Seoul	855	1	Male	1962	college	not marrie	have religion	999
140101	2014	Seoul	542	1	Male	1962	college	not marrie	not have religion	999
170101	2005	Seoul	651	1	Male	1964	Elementary	not marrie	not have religion	513
170101	2008	Seoul	1047	1	Male	1964	Elementary	not marrie	not have religion	530
170101	2009	Seoul	1102	1	Male	1964	Elementary	not marrie	not have religion	530
170101	2010	Seoul	1026	1	Male	1964	Elementary	not marrie	not have religion	530
170101	2011	Seoul	796	1	Male	1964	Elementary	not marrie	not have religion	530
170101	2012	Seoul	1032	1	Male	1964	Elementary	not marrie	not have religion	521
170101	2013	Seoul	1154	1	Male	1964	Elementary	not marrie	not have religion	530
170101	2014	Seoul	1363	1	Male	1964	Elementary	not marrie	not have religion	530
170101	2015	Seoul	1588	1	Male	1964	Elementary	not marrie	not have religion	530

3. Data folder: https://drive.google.com/drive/folders/1w39Oh5uCs1AHOW4AJcFpmqUu7lo2RdZw?usp=sharing