Unlocking Opportunities: A Data-Driven Path to Alleviate Unemployment in the Philippines

DATA SCIENTIST CAPSTONE COURSE

JUVEN DALE Q. COLASTE
PROJECT SPARTA SCHOLAR

INTRODUCTION

Background of the Study

The Philippines, a country brimming with potential and resilience, grapples with a significant challenge of unemployment. This issue isn't just about numbers; it's about the dreams and livelihoods of countless Filipinos.

The "PHL-PSA-LFS-2021-01-PUF" dataset is a unique resource, providing a comprehensive look at the Philippine labor force. It allows us to delve into the complexities of unemployment, dissect its roots, and imagine potential solutions.

With the data at our fingertips, we can unlock a treasure trove of insights into the nature and causes of unemployment, paving the way for a brighter future for the Philippines.

STATEMENT OF THE PROBLEM

Portraying the Unemployment Landscape

- What is the current unemployment rate in the Philippines, and how has it changed over time?
- What are the top primary occupations in need of assistance when people are searching for employment?

Diagnosing the Unemployment Ailments

- Why is there a variation in employment status among different demographic in the Philippines, and are there specific factors contributing to this discrepancy?
- Why do different age groups have their unique ways of job hunting, and can we figure out what's causing these differences?

Predicting the Future of Unemployment:

- Can we predict the likelihood of an individual in the Philippines being employed or unemployed, and what factors have the most significant impact on their employment status?

SCOPE & LIMITATIONS

- The analysis is based on the data available in the "PHL-PSA-LFS-2021-01-PUF" dataset. Any inaccuracies or limitations within this dataset may impact the accuracy of the findings.
- The dataset does not include information on all types of unemployment, such as disguised unemployment and frictional unemployment. This means that the unemployment rate calculated from the dataset may be underestimated.
- While the dataset can provide correlations between various factors and unemployment, it may not establish causation. Other hidden variables or external factors not covered in the dataset may influence unemployment trend.

- Predictions about future unemployment trends are based on historical data and assumptions. They do not account for unforeseen events or policy changes that may influence employment outcomes.

LITERATURE REVIEW

Available Research

Over the years, the Philippines has grappled with persistently high unemployment rates, which extend their impact beyond individuals and their families, with broader economic and social implications (Salvosa, 2015).

In this context, analytics and data science have emerged as increasingly vital tools in the battle against unemployment. These fields enable researchers to discern patterns within labor market data, make forecasts regarding employment trends, and offer evidence-based policy recommendations through the use of sophisticated analytical techniques (Smaldone et al., 2022).

Effective strategies aimed at reducing unemployment encompass initiatives such as skills development programs, job matching platforms, and targeted assistance for vulnerable populations. Leveraging data-driven insights becomes paramount in shaping the design and execution of these interventions to ensure they achieve the highest possible impact (Why Should We Integrate Income and Employment Support? A Conceptual and Empirical Investigation, n.d.).

Gap Analysis

A significant knowledge gap exists in the form of a limited body of research dedicated to unraveling the root causes of unemployment in the Philippines (Poverty in the Philippines: Causes, Constraints and Opportunities, 2022). This deficiency underscores a lack of comprehensive investigation and analysis into the intricate web of factors and circumstances that contribute to the country's elevated unemployment rates.

The absence of thorough research has implications for our ability to fully grasp the variations in unemployment rates among different geographic regions and demographic groups within the Philippines (Brooks & Author_Id, 2002). While age, gender, and geographical location may play roles in these disparities, the absence of comprehensive research makes it challenging to pinpoint the underlying causes.

PROPOSED RESEARCH DESIGN

Step-by-Step Design to Solve the Problem

- The researcher will begin by identifying vital dataset sources like government demographics and publicly available datasets such as labor force and household surveys, which will provide information on the Philippine labor force.
- The dataset will undergo thorough cleansing to eliminate errors or inconsistencies, integrate metadata, and reduce data anomalies.

- To gain insights and understand the challenges faced by unemployed Filipinos, descriptive analytics techniques like control charts and Pareto chart analysis will be employed.
- Diagnostic analytics methods, such as the Chi-square test and One-way ANOVA, will be utilized to precisely pinpoint the factors contributing to unemployment in the Philippines.
- Predictive analytics, facilitated by logistic regression, will be used to not only identify but also assess the factors influencing the likelihood of future unemployment.
- Lastly, this research will aim to investigate, propose, and advocate for policies and programs aimed at alleviating the unemployment issue in the Philippines.

Proposed Data Analysis Methodologies

Quantitative

This research will encompass three statistical methods: descriptive, diagnostic, and predictive statistics. Descriptive statistics will summarize and present data effectively, revealing key characteristics and trends. Diagnostic statistics, such as the Chi-square test and One-way ANOVA, will uncover associations and patterns in categorical variables related to employment. Predictive statistics, exemplified by logistic regression, will help anticipate future unemployment probabilities. Visual representations, like charts and graphs, will offer accessible data visualization.

Qualitative

In qualitative analysis, the data will play a crucial role in extracting meaningful insights through advanced analysis techniques to uncover key insights for unemployment strategy and challenges. Once associations are identified, the researcher will interpret their meaning and significance. This will entail a thorough exploration beyond surface-level analysis to fully grasp the implications of these themes within the context of the research goals.

IMPLEMENTATION & ANALYSES

Data Collection

Getting the right data for the capstone project was a headache. The researcher had decided to work on "unemployment" national issue, and by using data from the Philippines would be the key to finding a solution.

Data hunting and research was done on various government websites that lead to a whopping 100+ datasets collected. After sifting through all the collected information, the researcher finally settled on three (3) datasets you can see below.

LFS PUF January 2021.xlsx

Data Source: https://psada.psa.gov.ph

The "LFS PUF January 2021" is a quarterly survey of households in all parts of the country (in this case, 1st quarter – January 2021). It's the collected data on the demographic and socio-economic characteristics of all residents aged 15 and over, regardless of their sex, race, religion, citizenship, marital status, educational attainment, or economic status.

lfs_january_2021_metadata(dictionary).xlsx

Data Source: https://psada.psa.gov.ph

The "lfs_january_2021_metadata" dictionary holds information about the" LFS PUF January 2021". It has the details about the dataset and structure, making it easier to understand and work with the survey results.

phi-key-indicators-2023.xlsx

Data Source: https://www.adb.org/

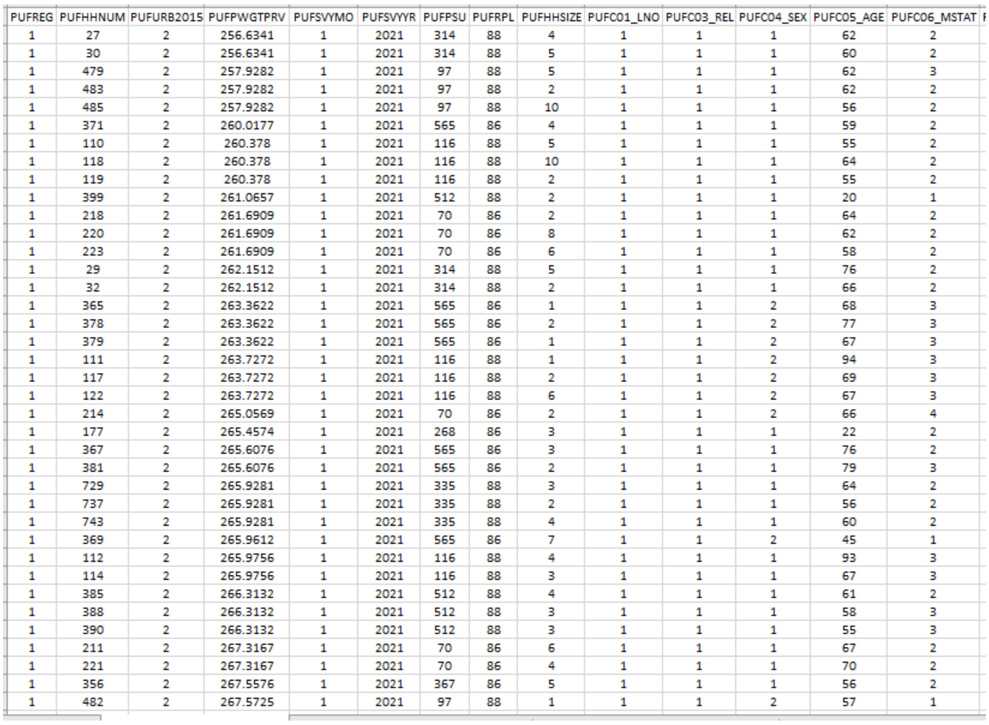
The "phi-key-indicators-2023" contains essential information and statistics about the Philippines from year 2000 to 2023. This data cover various key aspects such as the economy, population, and other important factors that provide an overview of the country's status in that year.

The “*LFS PUF January 2021*” has the heaviest file size having 105mb that took almost three (3) or four (4) minutes before excel can open it while the other files took seconds only, "*phi-key-indicators-2023*" has 130kb and *"lfs_january_2021_metadata"* with 71kb filesizes.

Data Exploration, Cleaning and Wrangling

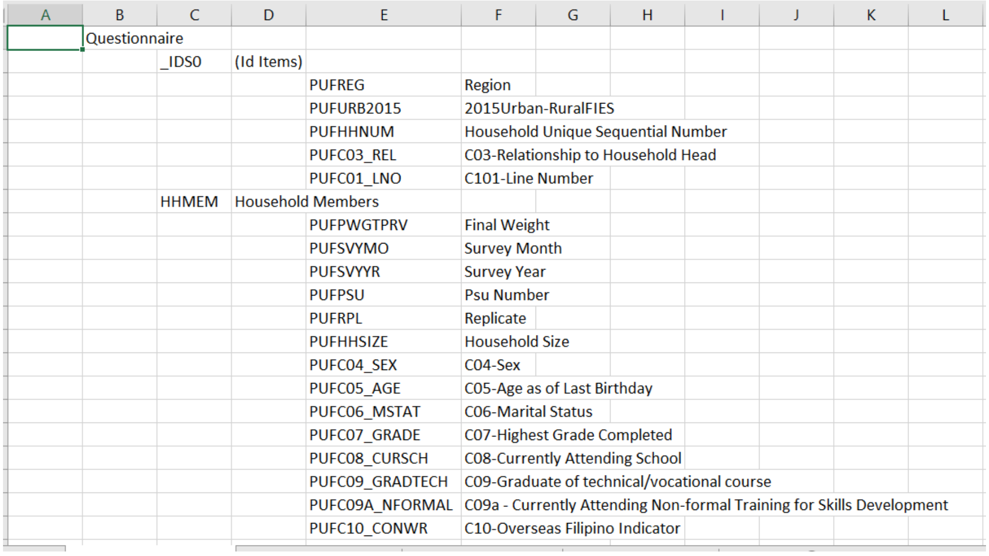
The table headers and column inputs from the survey of “*LFS PUF January 2021*” are in special codes and are to be decoded based in *"lfs_january_2021_metadata”* as shown below:

LFS PUF January 2021:



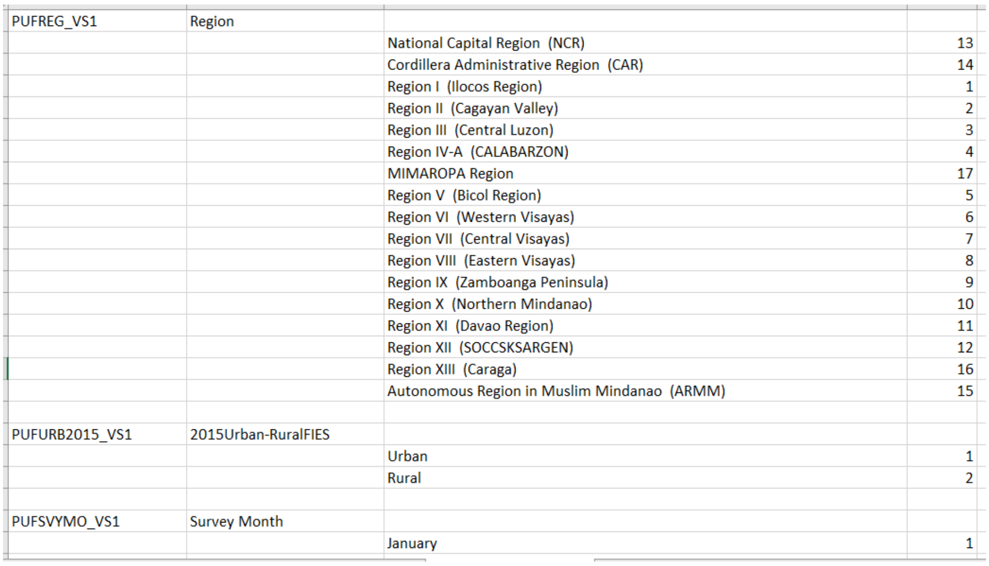
PUFREG	PUFHHNUM	PUFURB2015	PUFPWGTPRV	PUFSVYMO	PUFSVYYR	PUFPSU	PUFRPL	PUFHHSIZE	PUFC01_LNO	PUFC03_REL	PUFC04_SEX	PUFC05_AGE	PUFC06_MSTAT
1	27	2	256.6341	1	2021	314	88	4	1	1	1	62	2
1	30	2	256.6341	1	2021	314	88	5	1	1	1	60	2
1	479	2	257.9282	1	2021	97	88	5	1	1	1	62	3
1	483	2	257.9282	1	2021	97	88	2	1	1	1	62	2
1	485	2	257.9282	1	2021	97	88	10	1	1	1	56	2
1	371	2	260.0177	1	2021	565	86	4	1	1	1	59	2
1	110	2	260.378	1	2021	116	88	5	1	1	1	55	2
1	118	2	260.378	1	2021	116	88	10	1	1	1	64	2
1	119	2	260.378	1	2021	116	88	2	1	1	1	55	2
1	399	2	261.0657	1	2021	512	88	2	1	1	1	20	1
1	218	2	261.6909	1	2021	70	86	2	1	1	1	64	2
1	220	2	261.6909	1	2021	70	86	8	1	1	1	62	2
1	223	2	261.6909	1	2021	70	86	6	1	1	1	58	2
1	29	2	262.1512	1	2021	314	88	5	1	1	1	76	2
1	32	2	262.1512	1	2021	314	88	2	1	1	1	66	2
1	365	2	263.3622	1	2021	565	86	1	1	1	2	68	3
1	378	2	263.3622	1	2021	565	86	2	1	1	2	77	3
1	379	2	263.3622	1	2021	565	86	1	1	1	2	67	3
1	111	2	263.7272	1	2021	116	88	1	1	1	2	94	3
1	117	2	263.7272	1	2021	116	88	2	1	1	2	69	3
1	122	2	263.7272	1	2021	116	88	6	1	1	2	67	3
1	214	2	265.0569	1	2021	70	86	2	1	1	2	66	4
1	177	2	265.4574	1	2021	268	86	3	1	1	1	22	2
1	367	2	265.6076	1	2021	565	86	3	1	1	1	76	2
1	381	2	265.6076	1	2021	565	86	2	1	1	1	79	3
1	729	2	265.9281	1	2021	335	88	3	1	1	1	64	2
1	737	2	265.9281	1	2021	335	88	2	1	1	1	56	2
1	743	2	265.9281	1	2021	335	88	4	1	1	1	60	2
1	369	2	265.9612	1	2021	565	86	7	1	1	2	45	1
1	112	2	265.9756	1	2021	116	88	4	1	1	1	93	3
1	114	2	265.9756	1	2021	116	88	3	1	1	1	67	3
1	385	2	266.3132	1	2021	512	88	4	1	1	1	61	2
1	388	2	266.3132	1	2021	512	88	3	1	1	1	58	3
1	390	2	266.3132	1	2021	512	88	3	1	1	1	55	3
1	211	2	267.3167	1	2021	70	86	6	1	1	1	67	2
1	221	2	267.3167	1	2021	70	86	4	1	1	1	70	2
1	356	2	267.5576	1	2021	367	86	5	1	1	1	56	2
1	482	2	267.5725	1	2021	97	88	1	1	1	2	57	1

lfs_january_2021_metadata (column values):



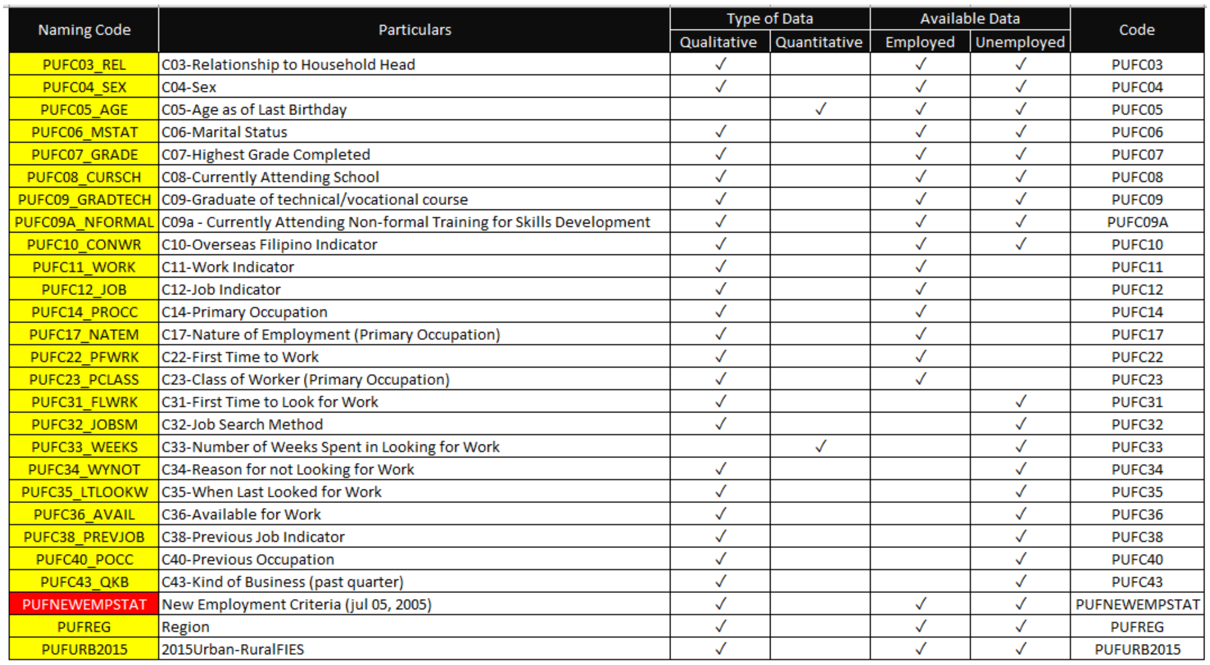
A	B	C	D	E	F	G	H	I	J	K	L
	Questionnaire										
		_IDS0	(Id Items)								
				PUFREG	Region						
				PUFURB2015	2015Urban-RuralFIES						
				PUFHHNUM	Household Unique Sequential Number						
				PUFC03_REL	C03-Relationship to Household Head						
				PUFC01_LNO	C101-Line Number						
		HHMEM	Household Members								
				PUFPWGTPRV	Final Weight						
				PUFSVYMO	Survey Month						
				PUFSVYYR	Survey Year						
				PUFPSU	Psu Number						
				PUFRPL	Replicate						
				PUFHHSIZE	Household Size						
				PUFC04_SEX	C04-Sex						
				PUFC05_AGE	C05-Age as of Last Birthday						
				PUFC06_MSTAT	C06-Marital Status						
				PUFC07_GRADE	C07-Highest Grade Completed						
				PUFC08_CURSCH	C08-Currently Attending School						
				PUFC09_GRADTECH	C09-Graduate of technical/vocational course						
				PUFC09A_NFORMAL	C09a - Currently Attending Non-formal Training for Skills Development						
				PUFC10_CONWR	C10-Overseas Filipino Indicator						

lfs_january_2021_metadata (input values):



PUFREG_VS1	Region		
		National Capital Region (NCR)	13
		Cordillera Administrative Region (CAR)	14
		Region I (Ilocos Region)	1
		Region II (Cagayan Valley)	2
		Region III (Central Luzon)	3
		Region IV-A (CALABARZON)	4
		MIMAROPA Region	17
		Region V (Bicol Region)	5
		Region VI (Western Visayas)	6
		Region VII (Central Visayas)	7
		Region VIII (Eastern Visayas)	8
		Region IX (Zamboanga Peninsula)	9
		Region X (Northern Mindanao)	10
		Region XI (Davao Region)	11
		Region XII (SOCCSKSARGEN)	12
		Region XIII (Caraga)	16
		Autonomous Region in Muslim Mindanao (ARMM)	15
PUFURB2015_VS1	2015Urban-RuralFIES		
		Urban	1
		Rural	2
PUFSVYMO_VS1	Survey Month		
		January	1

The dataset was cleaned by removing columns that are not needed in the study and left with the table below:

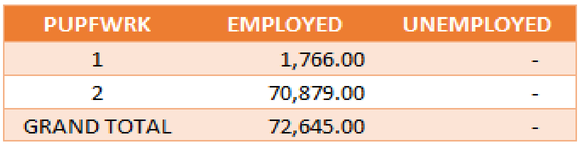


Naming Code	Particulars	Type of Data		Available Data		Code
		Qualitative	Quantitative	Employed	Unemployed	
PUFC03_REL	C03-Relationship to Household Head	✓		✓	✓	PUFC03
PUFC04_SEX	C04-Sex	✓		✓	✓	PUFC04
PUFC05_AGE	C05-Age as of Last Birthday		✓	✓	✓	PUFC05
PUFC06_MSTAT	C06-Marital Status	✓		✓	✓	PUFC06
PUFC07_GRADE	C07-Highest Grade Completed	✓		✓	✓	PUFC07
PUFC08_CURSCH	C08-Currently Attending School	✓		✓	✓	PUFC08
PUFC09_GRADTECH	C09-Graduate of technical/vocational course	✓		✓	✓	PUFC09
PUFC09A_NFORMAL	C09a - Currently Attending Non-formal Training for Skills Development	✓		✓	✓	PUFC09A
PUFC10_CONWR	C10-Overseas Filipino Indicator	✓		✓	✓	PUFC10
PUFC11_WORK	C11-Work Indicator	✓		✓		PUFC11
PUFC12_JOB	C12-Job Indicator	✓		✓		PUFC12
PUFC14_PROCC	C14-Primary Occupation	✓		✓		PUFC14
PUFC17_NATEM	C17-Nature of Employment (Primary Occupation)	✓		✓		PUFC17
PUFC22_PFWRK	C22-First Time to Work	✓		✓		PUFC22
PUFC23_PCLASS	C23-Class of Worker (Primary Occupation)	✓		✓		PUFC23
PUFC31_FLWRK	C31-First Time to Look for Work	✓			✓	PUFC31
PUFC32_JOBSM	C32-Job Search Method	✓			✓	PUFC32
PUFC33_WEEKS	C33-Number of Weeks Spent in Looking for Work		✓		✓	PUFC33
PUFC34_WYNOT	C34-Reason for not Looking for Work	✓			✓	PUFC34
PUFC35_LTLOOKW	C35-When Last Looked for Work	✓			✓	PUFC35
PUFC36_AVAIL	C36-Available for Work	✓			✓	PUFC36
PUFC38_PREVJOB	C38-Previous Job Indicator	✓			✓	PUFC38
PUFC40_POCC	C40-Previous Occupation	✓			✓	PUFC40
PUFC43_QKB	C43-Kind of Business (past quarter)	✓			✓	PUFC43
PUFNEWEMPSTAT	New Employment Criteria (jul 05, 2005)	✓		✓	✓	PUFNEWEMPSTAT
PUFREG	Region	✓		✓	✓	PUFREG
PUFURB2015	2015Urban-RuralFIES	✓		✓	✓	PUFURB2015

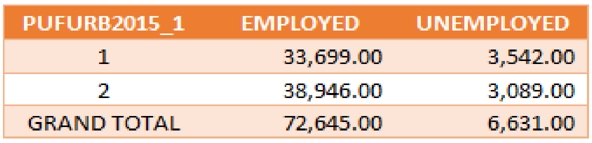
The first column (naming code) from above are the column names from our *"LFS PUF January 2021"* and the "red" shaded cell is the assumed categorical classification to be used in our predictive analysis modeling as one of the solutions in the researcher's statement of the problem.

Upon checking on *"LFS PUF January 2021",* the columns are mostly filled with numbers however upon referring them into our metadata, they are accounted either an ordinal or categorical. This means most of the data inputs are qualitative as shown in the above table under "Type of Table". Only "PUFC05_AGE" (C05-Age as of Last Birthday) and "PUFC33_WEEKS" (C33-Number of Weeks Spent in Looking for Work) are accounted as quantitative type.

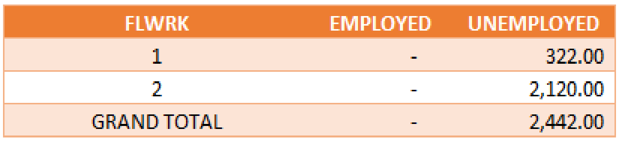
The purpose of the column header “Available Data” is to determine if the column headers from *“LFS PUF January 2021”* has data count input in terms of “Employed” and “Unemployed” as shown below:



PUPFWRK	EMPLOYED	UNEMPLOYED
1	1,766.00	-
2	70,879.00	-
GRAND TOTAL	72,645.00	-



PUFURB2015_1	EMPLOYED	UNEMPLOYED
1	33,699.00	3,542.00
2	38,946.00	3,089.00
GRAND TOTAL	72,645.00	6,631.00



FLWRK	EMPLOYED	UNEMPLOYED
1	-	322.00
2	-	2,120.00
GRAND TOTAL	-	2,442.00

The columns with no value count under “unemployed” or “employed” means that our dataset doesn’t contain participants that falls under that category. This also explains the limits from our scope and limitations.

In other words, it means that our dataset may not encompass the full spectrum of employment statuses, and certain categories, such as "unemployed" or "employed," might not be adequately represented due to the dataset's inherent constraints. This limitation should be taken into account when interpreting and drawing conclusions from the data analysis.

This also means that some of the columns may not be used depending on the requirements of the specific analytical technique to be used in the research.

Descriptive Analytics

The researcher used descriptive analytics, including control chart and Pareto chart analysis, to gain insights and understand the challenges faced by unemployed Filipinos.

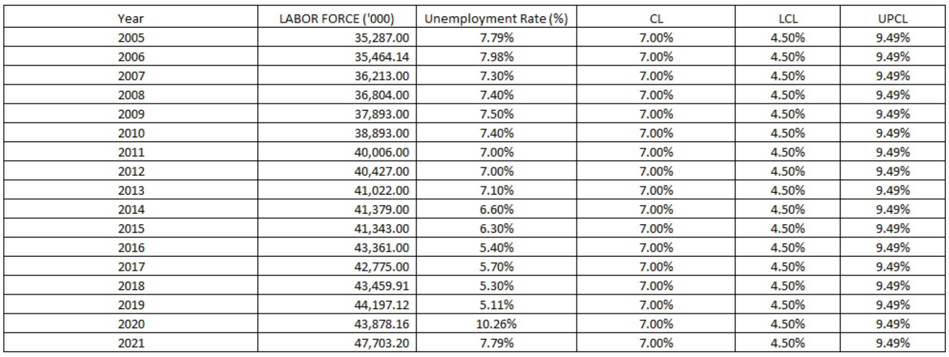
Control Chart Analysis

Using Control Chart Analysis, the researcher identified unusual employment rate patterns over time. Data from "phi-key-indicators-2023," including "Year," "Labor Force," and unemployment rate, was used for this analysis. To create the Control Chart, the researcher calculated the "Control Limit," "Lower Control Limit," and "Upper Control Limit."

$$UCL = \overline{X} + E_2\overline{R}$$

$$CL = \overline{X} = \sum_{i=1,k} X_i \Big/ k$$

$$LCL = \overline{X} - E_2\overline{R}$$



Year	LABOR FORCE ('000)	Unemployment Rate (%)	CL	LCL	UPCL
2005	35,287.00	7.79%	7.00%	4.50%	9.49%
2006	35,464.14	7.98%	7.00%	4.50%	9.49%
2007	36,213.00	7.30%	7.00%	4.50%	9.49%
2008	36,804.00	7.40%	7.00%	4.50%	9.49%
2009	37,893.00	7.50%	7.00%	4.50%	9.49%
2010	38,893.00	7.40%	7.00%	4.50%	9.49%
2011	40,006.00	7.00%	7.00%	4.50%	9.49%
2012	40,427.00	7.00%	7.00%	4.50%	9.49%
2013	41,022.00	7.10%	7.00%	4.50%	9.49%
2014	41,379.00	6.60%	7.00%	4.50%	9.49%
2015	41,343.00	6.30%	7.00%	4.50%	9.49%
2016	43,361.00	5.40%	7.00%	4.50%	9.49%
2017	42,775.00	5.70%	7.00%	4.50%	9.49%
2018	43,459.91	5.30%	7.00%	4.50%	9.49%
2019	44,197.12	5.11%	7.00%	4.50%	9.49%
2020	43,878.16	10.26%	7.00%	4.50%	9.49%
2021	47,703.20	7.79%	7.00%	4.50%	9.49%

The table output (above) was produced by using the formula for UCL, CL & LCL as "Upper Control Limit", "Control Limit" and "Lower Control Limit", respectively.

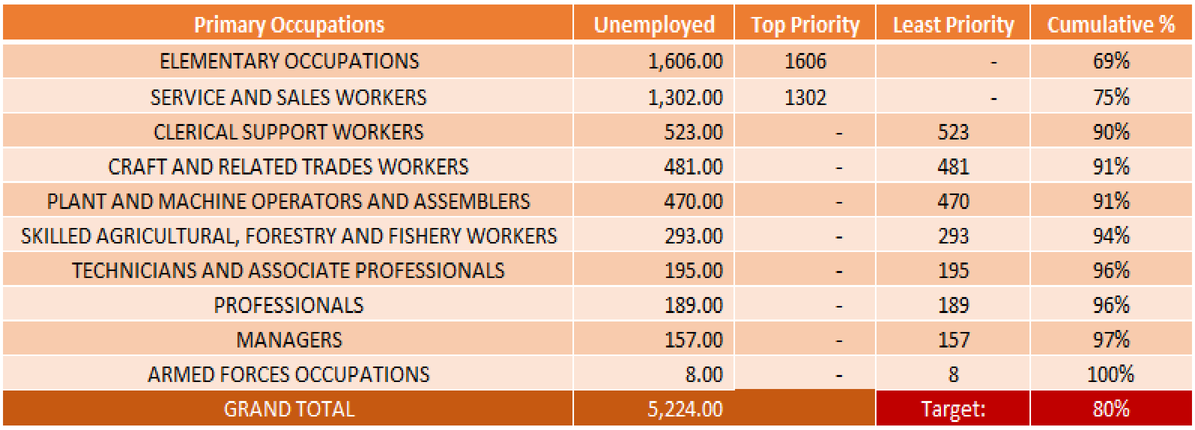
The Center Line (CL) is calculated as the average or mean of the data points in the dataset. In this case, it represents the average unemployment rate over a certain time period, such as the years from 2005 to 2021.

The Lower Control Limit (LCL) and Upper Control Limit (UPCL) are calculated based on the standard deviation parameters of the dataset.

The CL at 7.00%, indicating the average unemployment rate, while the LCL and UPCL has values of 4.50% and 9.49%. If the unemployment rate falls between the LCL and UPCL, it is considered within control; if it falls outside these limits, it can be flagged as a significant deviation from the expected or normal range. These control limits help in identifying periods or years where the unemployment rate deviates from the expected range.

Pareto Chart Analysis

The Pareto Chart will identify and prioritize the most impactful factors contributing to unemployment, allowing us to focus on these critical. The advanced descriptive analytical tool was used to determine the primary occupations that are unemployed that needs assistance to find jobs with the aim of greatly reducing unemployment by 80% as shown below:



Primary Occupations	Unemployed	Top Priority	Least Priority	Cumulative %
ELEMENTARY OCCUPATIONS	1,606.00	1606	-	69%
SERVICE AND SALES WORKERS	1,302.00	1302	-	75%
CLERICAL SUPPORT WORKERS	523.00	-	523	90%
CRAFT AND RELATED TRADES WORKERS	481.00	-	481	91%
PLANT AND MACHINE OPERATORS AND ASSEMBLERS	470.00	-	470	91%
SKILLED AGRICULTURAL, FORESTRY AND FISHERY WORKERS	293.00	-	293	94%
TECHNICIANS AND ASSOCIATE PROFESSIONALS	195.00	-	195	96%
PROFESSIONALS	189.00	-	189	96%
MANAGERS	157.00	-	157	97%
ARMED FORCES OCCUPATIONS	8.00	-	8	100%
GRAND TOTAL	5,224.00		Target:	80%

The table is from the "LFS PUF January 2021" dataset. It includes columns for primary occupations, the number of unemployed individuals in each occupation, the top priority occupations in need of assistance, the least priority occupations in need of assistance, and the cumulative percentage having a target of 80% of unemployment reduction.

To maximize the use of cumulative percentage, the table output was arranged in decreasing order based on the number of unemployed persons per primary occupation. As a result, when we reach the 80% target, primary occupations with a cumulative percentage below 80% (starting from the top) will be considered top priority, while the rest will fall into the least priority category.

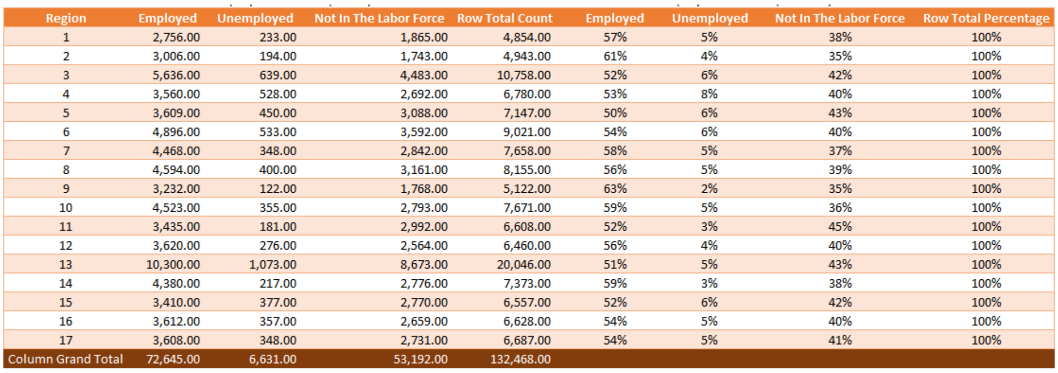
Diagnostic Analytics

Chi-square Test of Independence

The Chi-square Test of Independence is a statistical test used to determine if there is a significant association or relationship between two categorical variables. This advanced diagnostic analytic technique was used to determine the strength of association from employment status to region demographic profile, highest grade completed and technical/vocation graduate columns from the dataset "LFS PUF January 2021" having the pivot and association table shown below:

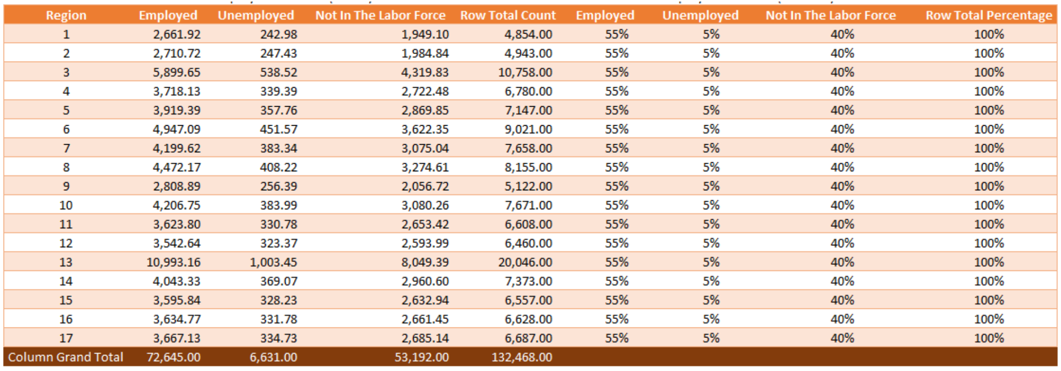
REGION DEMOGRAPHIC PROFILE

Pivot Table



Region	Employed	Unemployed	Not In The Labor Force	Row Total Count	Employed	Unemployed	Not In The Labor Force	Row Total Percentage
1	2,756.00	233.00	1,865.00	4,854.00	57%	5%	38%	100%
2	3,006.00	194.00	1,743.00	4,943.00	61%	4%	35%	100%
3	5,636.00	639.00	4,483.00	10,758.00	52%	6%	42%	100%
4	3,560.00	528.00	2,692.00	6,780.00	53%	8%	40%	100%
5	3,609.00	450.00	3,088.00	7,147.00	50%	6%	43%	100%
6	4,896.00	533.00	3,592.00	9,021.00	54%	6%	40%	100%
7	4,468.00	348.00	2,842.00	7,658.00	58%	5%	37%	100%
8	4,594.00	400.00	3,161.00	8,155.00	56%	5%	39%	100%
9	3,232.00	122.00	1,768.00	5,122.00	63%	2%	35%	100%
10	4,523.00	355.00	2,793.00	7,671.00	59%	5%	36%	100%
11	3,435.00	181.00	2,992.00	6,608.00	52%	3%	45%	100%
12	3,620.00	276.00	2,564.00	6,460.00	56%	4%	40%	100%
13	10,300.00	1,073.00	8,673.00	20,046.00	51%	5%	43%	100%
14	4,380.00	217.00	2,776.00	7,373.00	59%	3%	38%	100%
15	3,410.00	377.00	2,770.00	6,557.00	52%	6%	42%	100%
16	3,612.00	357.00	2,659.00	6,628.00	54%	5%	40%	100%
17	3,608.00	348.00	2,731.00	6,687.00	54%	5%	41%	100%
Column Grand Total	72,645.00	6,631.00	53,192.00	132,468.00				

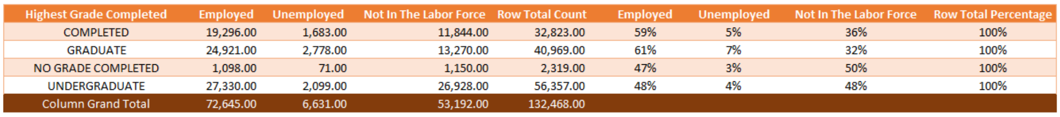
Association Table



Region	Employed	Unemployed	Not In The Labor Force	Row Total Count	Employed	Unemployed	Not In The Labor Force	Row Total Percentage
1	2,661.92	242.98	1,949.10	4,854.00	55%	5%	40%	100%
2	2,710.72	247.43	1,984.84	4,943.00	55%	5%	40%	100%
3	5,899.65	538.52	4,319.83	10,758.00	55%	5%	40%	100%
4	3,718.13	339.39	2,722.48	6,780.00	55%	5%	40%	100%
5	3,919.39	357.76	2,869.85	7,147.00	55%	5%	40%	100%
6	4,947.09	451.57	3,622.35	9,021.00	55%	5%	40%	100%
7	4,199.62	383.34	3,075.04	7,658.00	55%	5%	40%	100%
8	4,472.17	408.22	3,274.61	8,155.00	55%	5%	40%	100%
9	2,808.89	256.39	2,056.72	5,122.00	55%	5%	40%	100%
10	4,206.75	383.99	3,080.26	7,671.00	55%	5%	40%	100%
11	3,623.80	330.78	2,653.42	6,608.00	55%	5%	40%	100%
12	3,542.64	323.37	2,593.99	6,460.00	55%	5%	40%	100%
13	10,993.16	1,003.45	8,049.39	20,046.00	55%	5%	40%	100%
14	4,043.33	369.07	2,960.60	7,373.00	55%	5%	40%	100%
15	3,595.84	328.23	2,632.94	6,557.00	55%	5%	40%	100%
16	3,634.77	331.78	2,661.45	6,628.00	55%	5%	40%	100%
17	3,667.13	334.73	2,685.14	6,687.00	55%	5%	40%	100%
Column Grand Total	72,645.00	6,631.00	53,192.00	132,468.00				

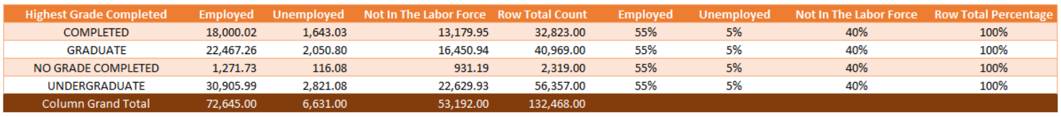
HIGHEST GRADE COMPLETED

Pivot Table



Highest Grade Completed	Employed	Unemployed	Not In The Labor Force	Row Total Count	Employed	Unemployed	Not In The Labor Force	Row Total Percentage
COMPLETED	19,296.00	1,683.00	11,844.00	32,823.00	59%	5%	36%	100%
GRADUATE	24,921.00	2,778.00	13,270.00	40,969.00	61%	7%	32%	100%
NO GRADE COMPLETED	1,098.00	71.00	1,150.00	2,319.00	47%	3%	50%	100%
UNDERGRADUATE	27,330.00	2,099.00	26,928.00	56,357.00	48%	4%	48%	100%
Column Grand Total	72,645.00	6,631.00	53,192.00	132,468.00				

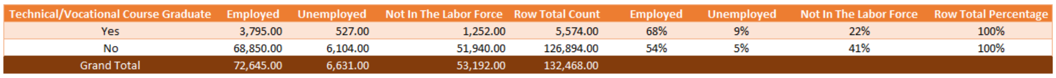
Association Table



Highest Grade Completed	Employed	Unemployed	Not In The Labor Force	Row Total Count	Employed	Unemployed	Not In The Labor Force	Row Total Percentage
COMPLETED	18,000.02	1,643.03	13,179.95	32,823.00	55%	5%	40%	100%
GRADUATE	22,467.26	2,050.80	16,450.94	40,969.00	55%	5%	40%	100%
NO GRADE COMPLETED	1,271.73	116.08	931.19	2,319.00	55%	5%	40%	100%
UNDERGRADUATE	30,905.99	2,821.08	22,629.93	56,357.00	55%	5%	40%	100%
Column Grand Total	72,645.00	6,631.00	53,192.00	132,468.00				

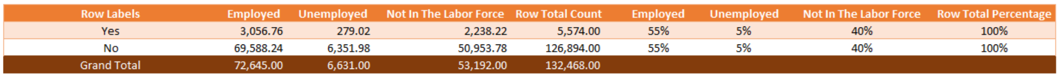
TECHNICAL VOCATIONAL GRADUATE

Pivot Table



Technical/Vocational Course Graduate	Employed	Unemployed	Not In The Labor Force	Row Total Count	Employed	Unemployed	Not In The Labor Force	Row Total Percentage
Yes	3,795.00	527.00	1,252.00	5,574.00	68%	9%	22%	100%
No	68,850.00	6,104.00	51,940.00	126,894.00	54%	5%	41%	100%
Grand Total	72,645.00	6,631.00	53,192.00	132,468.00				

Association Table



Row Labels	Employed	Unemployed	Not In The Labor Force	Row Total Count	Employed	Unemployed	Not In The Labor Force	Row Total Percentage
Yes	3,056.76	279.02	2,238.22	5,574.00	55%	5%	40%	100%
No	69,588.24	6,351.98	50,953.78	126,894.00	55%	5%	40%	100%
Grand Total	72,645.00	6,631.00	53,192.00	132,468.00				

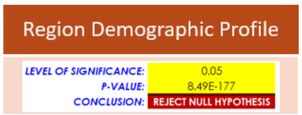
To create the association table for this test, the researcher referred to pivot tables, which provided a breakdown of percentage counts across "employed," "unemployed," and "not in the labor force" in relation to "region demographic profile," "highest grade completed," and "technical/vocational graduate."

The researcher then redistributed these percentage counts to construct an association table aimed at understanding the connections between these variables.

The researcher then assessed the strength of the relationship by using the Chi-square Test of Independence, which offers a P-value and its corresponding level of significance.

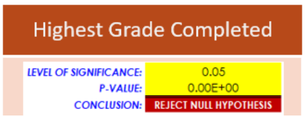
This information can help in making a decision regarding whether to accept or reject the null hypothesis, in which suggests no significant association or relationship between the categorical variables under investigation as shown below:

OVERALL TEST OF INDEPENDENCE OUTPUT



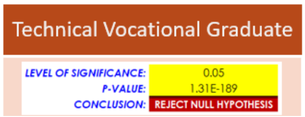
Region Demographic Profile

LEVEL OF SIGNIFICANCE:	0.05
P-VALUE:	8.49E-177
CONCLUSION:	REJECT NULL HYPOTHESIS



Highest Grade Completed

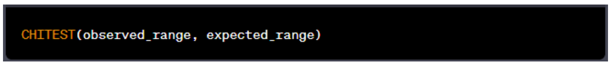
LEVEL OF SIGNIFICANCE:	0.05
P-VALUE:	0.00E+00
CONCLUSION:	REJECT NULL HYPOTHESIS



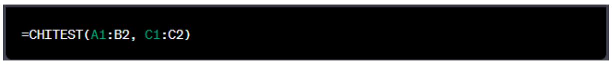
Technical Vocational Graduate

LEVEL OF SIGNIFICANCE:	0.05
P-VALUE:	1.31E-189
CONCLUSION:	REJECT NULL HYPOTHESIS

The outputs (p-value) were generated using the Excel data analysis function as shown below:



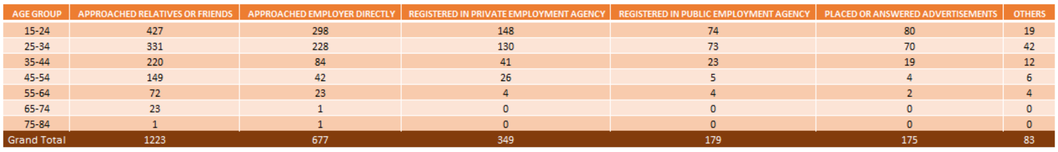
```
CHITEST(observed_range, expected_range)
```



```
=CHITEST(A1:B2, C1:C2)
```

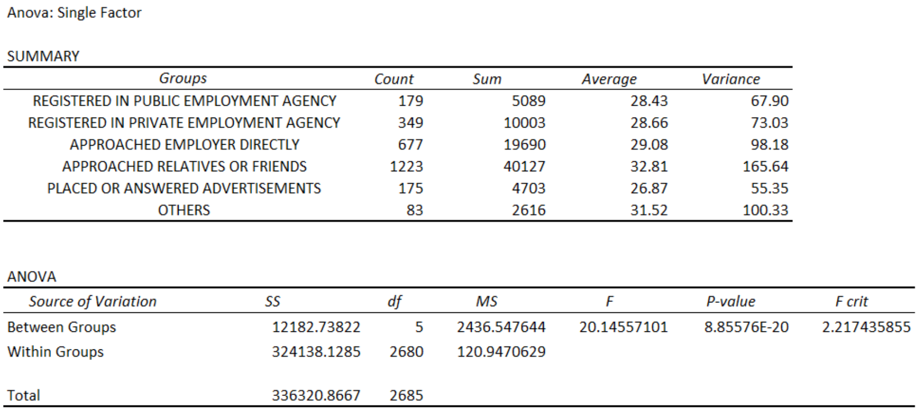
One Way Analysis of Variation (ANOVA)

The One-way ANOVA will assess the impact of factors affecting different groups approach job searching. The dataset was taken from "LFS PUF January 2021" having the overall age group table below:



AGE GROUP	APPROACHED RELATIVES OR FRIENDS	APPROACHED EMPLOYER DIRECTLY	REGISTERED IN PRIVATE EMPLOYMENT AGENCY	REGISTERED IN PUBLIC EMPLOYMENT AGENCY	PLACED OR ANSWERED ADVERTISEMENTS	OTHERS
15-24	427	298	148	74	80	19
25-34	331	228	130	73	70	42
35-44	220	84	41	23	19	12
45-54	149	42	26	5	4	6
55-64	72	23	4	4	2	4
65-74	23	1	0	0	0	0
75-84	1	1	0	0	0	0
Grand Total	1223	677	349	179	175	83

Please note that the table was for presentation purpose only and the input data that was used for the "One-way ANOVA" is the cleaned data wherein the age was not grouped yet. Knowing this, the output of the ANOVA is shown below:



Anova: Single Factor

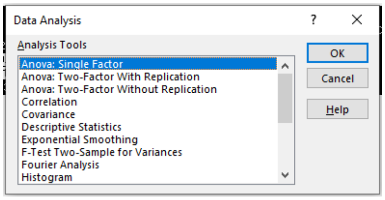
SUMMARY

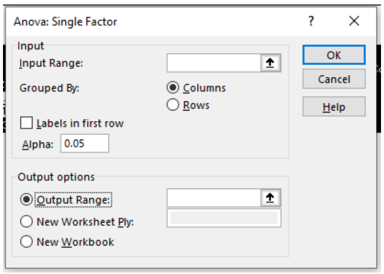
Groups	*Count*	*Sum*	*Average*	*Variance*
REGISTERED IN PUBLIC EMPLOYMENT AGENCY	179	5089	28.43	67.90
REGISTERED IN PRIVATE EMPLOYMENT AGENCY	349	10003	28.66	73.03
APPROACHED EMPLOYER DIRECTLY	677	19690	29.08	98.18
APPROACHED RELATIVES OR FRIENDS	1223	40127	32.81	165.64
PLACED OR ANSWERED ADVERTISEMENTS	175	4703	26.87	55.35
OTHERS	83	2616	31.52	100.33

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	12182.73822	5	2436.547644	20.14557101	8.85576E-20	2.217435855
Within Groups	324138.1285	2680	120.9470629			
Total	336320.8667	2685				

The output was taken by using excel data analysis function.





Grouped By: ● Columns

□ Labels in first row

Alpha: 0.05

● Output Range:

Predictive Analytics

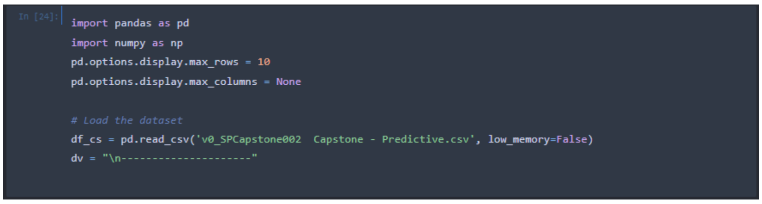
Logistic Regression

Logistic regression analysis is a statistical method used to model and analyze the relationship between a binary dependent variable (one that has only two possible outcomes, often coded as 0 and 1) and one or more independent variables (predictors or features). The primary goal of logistic regression is to predict the probability of the binary outcome – in our case, whether a person will find employment or not.

However, Logistic Regression was used in Jupyter Notebook rather than Excel in which needed to align with machine learning standards.

The researcher conducted additional comprehensive cleaning and filtering on the dataset "LFS PUF January 2021" to identify usable column features. This was done to reduce the amount of code needed in Jupyter Notebook. The resulting cleaned and filtered dataset was exported as "v0_SPCapstone002 Capstone – Predictive.csv" and will be imported for use in Jupyter Notebook, as demonstrated below:

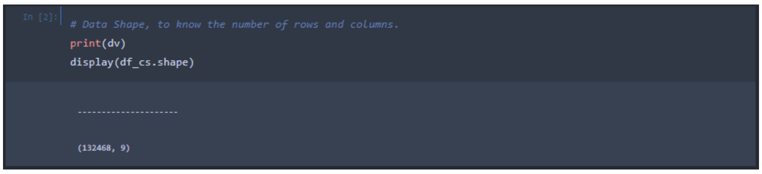
Dataset Loading/Importing



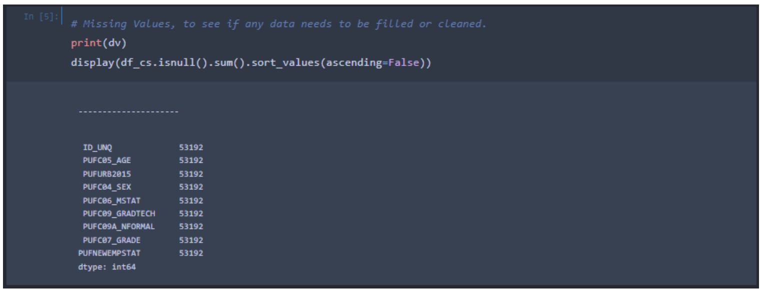
```
import pandas as pd
import numpy as np
pd.options.display.max_rows = 10
pd.options.display.max_columns = None

# Load the dataset
df_cs = pd.read_csv('v0_SPCapstone002  Capstone - Predictive.csv', low_memory=False)
dv = "\n---------------------"
```

After importing of data, the following are conducted accordingly:

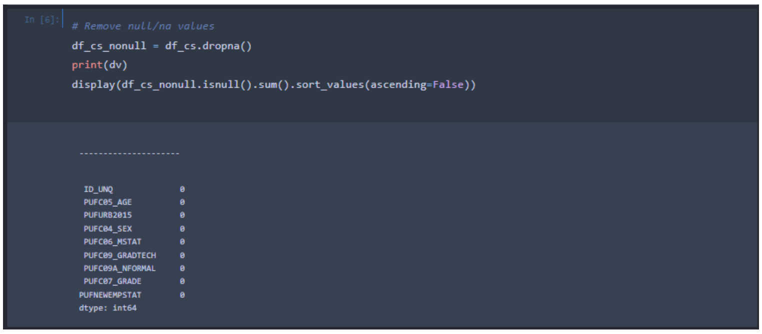
Exploratory Analysis



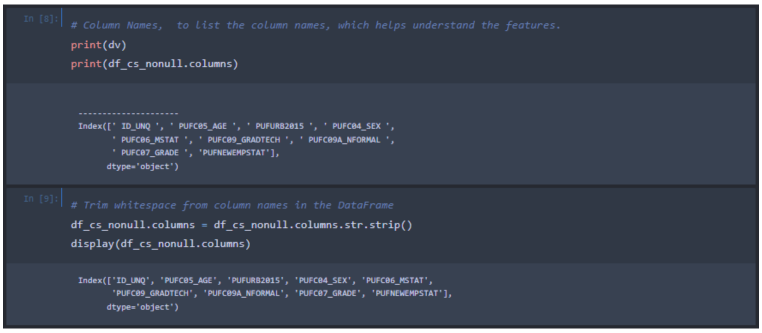
```
In [2]: # Data Shape, to know the number of rows and columns.
        print(dv)
        display(df_cs.shape)

         --------------------

         (132468, 9)
```



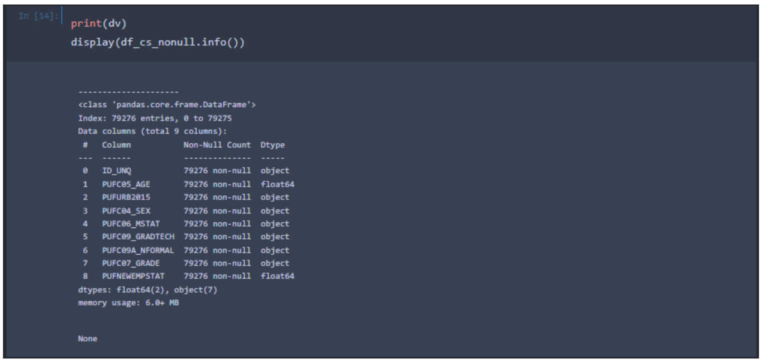
```
In [5]: # Missing Values, to see if any data needs to be filled or cleaned.
        print(dv)
        display(df_cs.isnull().sum().sort_values(ascending=False))

         --------------------

          ID_UNQ              53192
          PUFC05_AGE          53192
          PUFURB2015          53192
          PUFC04_SEX          53192
          PUFC06_MSTAT        53192
          PUFC09_GRADTECH     53192
          PUFC09A_NFORMAL     53192
          PUFC07_GRADE        53192
         PUFNEWEMPSTAT        53192
         dtype: int64
```



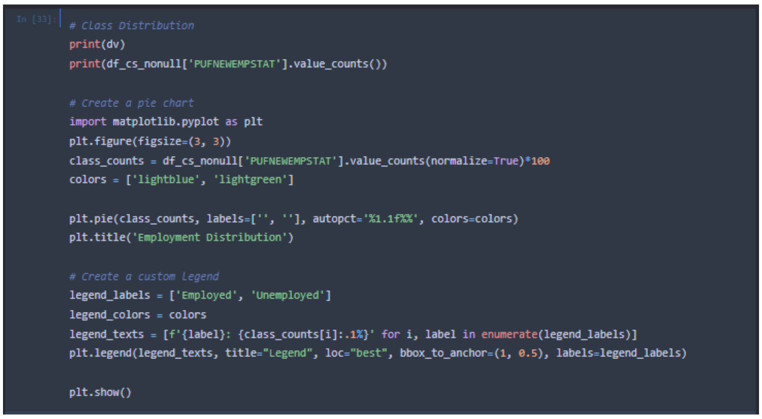
```
In [6]: # Remove null/na values
        df_cs_nonull = df_cs.dropna()
        print(dv)
        display(df_cs_nonull.isnull().sum().sort_values(ascending=False))

         --------------------

          ID_UNQ              0
          PUFC05_AGE          0
          PUFURB2015          0
          PUFC04_SEX          0
          PUFC06_MSTAT        0
          PUFC09_GRADTECH     0
          PUFC09A_NFORMAL     0
          PUFC07_GRADE        0
         PUFNEWEMPSTAT        0
         dtype: int64
```



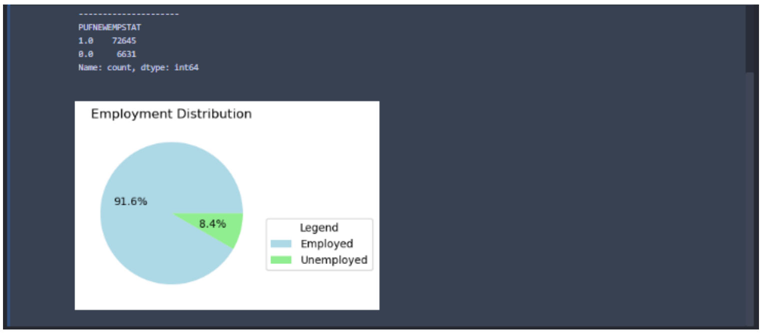
```
In [8]: # Column Names,  to list the column names, which helps understand the features.
        print(dv)
        print(df_cs_nonull.columns)

         --------------------
         Index([' ID_UNQ ', ' PUFC05_AGE ', ' PUFURB2015 ', ' PUFC04_SEX ',
                ' PUFC06_MSTAT ', ' PUFC09_GRADTECH ', ' PUFC09A_NFORMAL ',
                ' PUFC07_GRADE ', 'PUFNEWEMPSTAT'],
               dtype='object')

In [9]: # Trim whitespace from column names in the DataFrame
        df_cs_nonull.columns = df_cs_nonull.columns.str.strip()
        display(df_cs_nonull.columns)

         Index(['ID_UNQ', 'PUFC05_AGE', 'PUFURB2015', 'PUFC04_SEX', 'PUFC06_MSTAT',
                'PUFC09_GRADTECH', 'PUFC09A_NFORMAL', 'PUFC07_GRADE', 'PUFNEWEMPSTAT'],
               dtype='object')
```



```
In [14]: print(dv)
         display(df_cs_nonull.info())

         --------------------
         <class 'pandas.core.frame.DataFrame'>
         Index: 79276 entries, 0 to 79275
         Data columns (total 9 columns):
          #   Column           Non-Null Count  Dtype
         ---  ------           --------------  -----
          0   ID_UNQ           79276 non-null  object
          1   PUFC05_AGE       79276 non-null  float64
          2   PUFURB2015       79276 non-null  object
          3   PUFC04_SEX       79276 non-null  object
          4   PUFC06_MSTAT     79276 non-null  object
          5   PUFC09_GRADTECH  79276 non-null  object
          6   PUFC09A_NFORMAL  79276 non-null  object
          7   PUFC07_GRADE     79276 non-null  object
          8   PUFNEWEMPSTAT    79276 non-null  float64
         dtypes: float64(2), object(7)
         memory usage: 6.0+ MB


         None
```



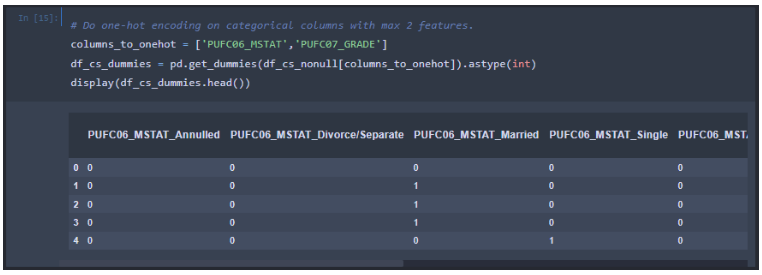
```
In [33]: # Class Distribution
print(dv)
print(df_cs_nonull['PUFNEWEMPSTAT'].value_counts())

# Create a pie chart
import matplotlib.pyplot as plt
plt.figure(figsize=(3, 3))
class_counts = df_cs_nonull['PUFNEWEMPSTAT'].value_counts(normalize=True)*100
colors = ['lightblue', 'lightgreen']

plt.pie(class_counts, labels=['', ''], autopct='%1.1f%%', colors=colors)
plt.title('Employment Distribution')

# Create a custom Legend
legend_labels = ['Employed', 'Unemployed']
legend_colors = colors
legend_texts = [f'{label}: {class_counts[i]:.1%}' for i, label in enumerate(legend_labels)]
plt.legend(legend_texts, title="Legend", loc="best", bbox_to_anchor=(1, 0.5), labels=legend_labels)

plt.show()
```



```
--------------------
PUFNEWEMPSTAT
1.0    72645
0.0     6631
Name: count, dtype: int64
```

Exploratory analysis was conducted to gain a deeper understanding of the dataset and its underlying characteristics, including dimensional assessment, scrutiny of null values, assessment of column data types, class distribution on binary output and evaluation of column names. This process provided a comprehensive foundation for subsequent data analysis, ensuring data completeness and reliability for advanced statistical and machine learning techniques.

Preprocessing: One-hot Encoding

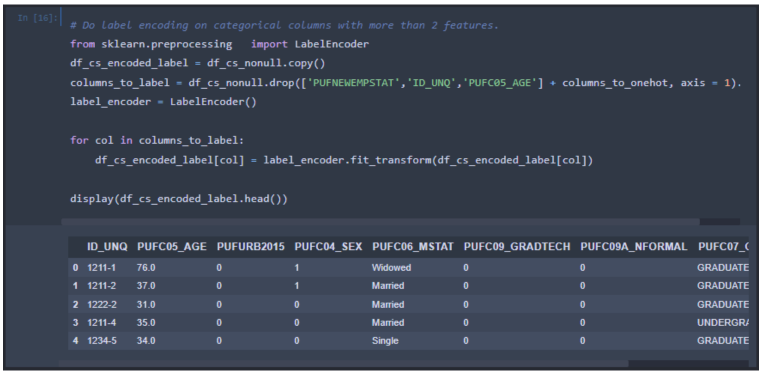


```
In [15]: # Do one-hot encoding on categorical columns with max 2 features.
columns_to_onehot = ['PUFC06_MSTAT','PUFC07_GRADE']
df_cs_dummies = pd.get_dummies(df_cs_nonull[columns_to_onehot]).astype(int)
display(df_cs_dummies.head())
```

	PUFC06_MSTAT_Annulled	PUFC06_MSTAT_Divorce/Separate	PUFC06_MSTAT_Married	PUFC06_MSTAT_Single	PUFC06_MST.
0	0	0	0	0	0
1	0	0	1	0	0
2	0	0	1	0	0
3	0	0	1	0	0
4	0	0	0	1	0

One-hot encoding allows us to represent categorical variables in a way that machine learning algorithms can understand and process effectively. The researcher applied one-hot encoding to the columns 'PUFC06_MSTAT' and 'PUFC07_GRADE' because each of these columns contains two (2) distinct features. This step was taken to prepare the data for machine learning analysis, where these encoded features will play a crucial role in our predictive model.

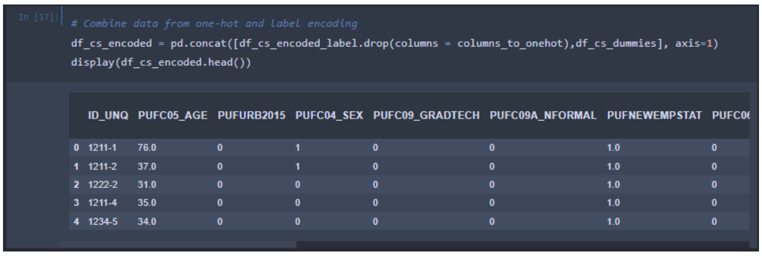
Preprocessing: Label Encoding



In [16]:
```
# Do label encoding on categorical columns with more than 2 features.
from sklearn.preprocessing   import LabelEncoder
df_cs_encoded_label = df_cs_nonull.copy()
columns_to_label = df_cs_nonull.drop(['PUFNEWEMPSTAT','ID_UNQ','PUFC05_AGE'] + columns_to_onehot, axis = 1).
label_encoder = LabelEncoder()

for col in columns_to_label:
    df_cs_encoded_label[col] = label_encoder.fit_transform(df_cs_encoded_label[col])

display(df_cs_encoded_label.head())
```

	ID_UNQ	PUFC05_AGE	PUFURB2015	PUFC04_SEX	PUFC06_MSTAT	PUFC09_GRADTECH	PUFC09A_NFORMAL	PUFC07_
0	1211-1	76.0	0	1	Widowed	0	0	GRADUATE
1	1211-2	37.0	0	1	Married	0	0	GRADUATE
2	1222-2	31.0	0	0	Married	0	0	GRADUATE
3	1211-4	35.0	0	0	Married	0	0	UNDERGRA
4	1234-5	34.0	0	0	Single	0	0	GRADUATE

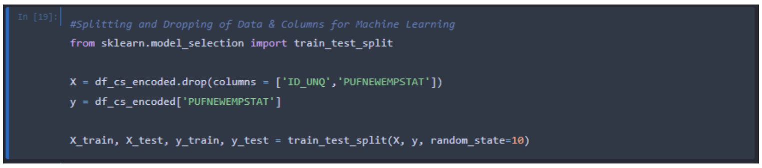
In this phase, label encoding was applied to columns with more than two (2) features. The purpose of label encoding is to enhance the machine learning process by transforming categorical data into a numerical format, making it more amenable to algorithmic calculations and predictions. This preparation step plays a crucial role in ensuring that the dataset is compatible with machine learning models, ultimately improving the accuracy and effectiveness of the analysis.



```
In [17]:  # Combine data from one-hot and label encoding
          df_cs_encoded = pd.concat([df_cs_encoded_label.drop(columns = columns_to_onehot),df_cs_dummies], axis=1)
          display(df_cs_encoded.head())
```

	ID_UNQ	PUFC05_AGE	PUFURB2015	PUFC04_SEX	PUFC09_GRADTECH	PUFC09A_NFORMAL	PUFNEWEMPSTAT	PUFC0
0	1211-1	76.0	0	1	0	0	1.0	0
1	1211-2	37.0	0	1	0	0	1.0	0
2	1222-2	31.0	0	0	0	0	1.0	0
3	1211-4	35.0	0	0	0	0	1.0	0
4	1234-5	34.0	0	0	0	0	1.0	0

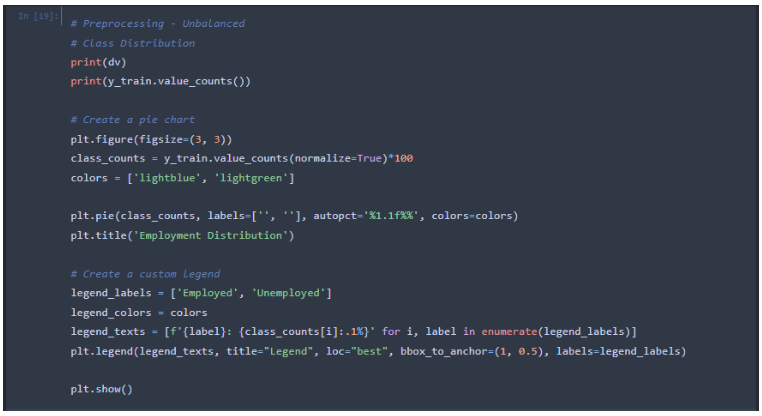
Preprocessing: Data Splitting (Train and Test)



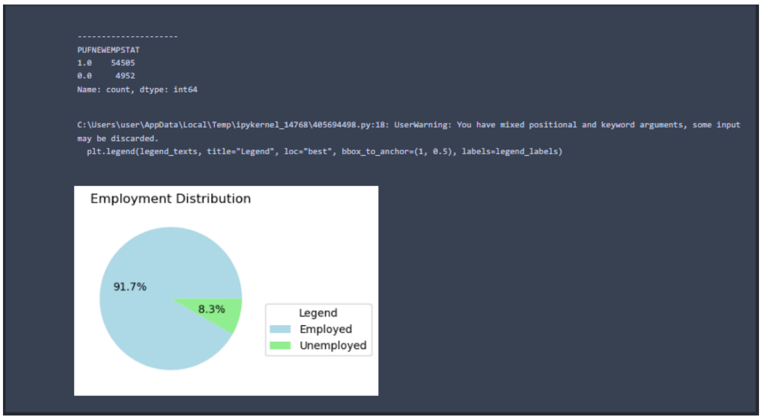
```
In [19]:  #Splitting and Dropping of Data & Columns for Machine Learning
          from sklearn.model_selection import train_test_split

          X = df_cs_encoded.drop(columns = ['ID_UNQ','PUFNEWEMPSTAT'])
          y = df_cs_encoded['PUFNEWEMPSTAT']

          X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=10)
```

The researcher split the data in advance in order to have data exclusivity. This will allow us to have clearer data processing across the training and testing datasets and increase our level of caution during further preprocessing.

Preprocessing: Unbalanced SMOTE

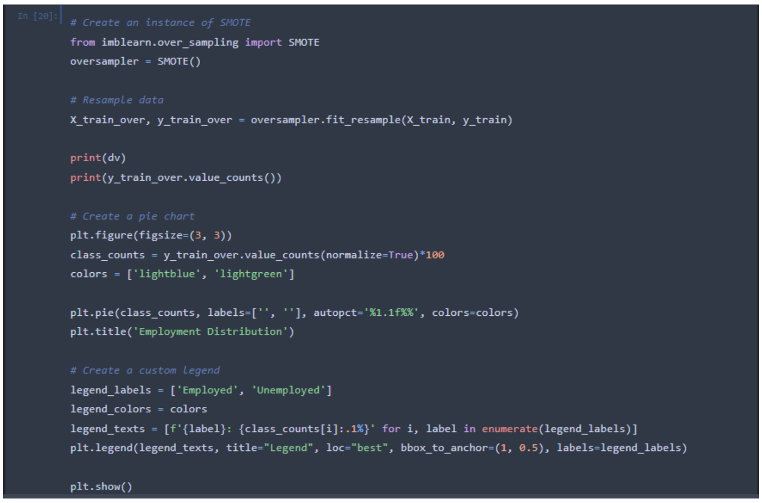


```
In [19]:  # Preprocessing - Unbalanced
          # Class Distribution
          print(dv)
          print(y_train.value_counts())

          # Create a pie chart
          plt.figure(figsize=(3, 3))
          class_counts = y_train.value_counts(normalize=True)*100
          colors = ['lightblue', 'lightgreen']

          plt.pie(class_counts, labels=['', ''], autopct='%1.1f%%', colors=colors)
          plt.title('Employment Distribution')

          # Create a custom legend
          legend_labels = ['Employed', 'Unemployed']
          legend_colors = colors
          legend_texts = [f'{label}: {class_counts[i]:.1%}' for i, label in enumerate(legend_labels)]
          plt.legend(legend_texts, title="Legend", loc="best", bbox_to_anchor=(1, 0.5), labels=legend_labels)

          plt.show()
```

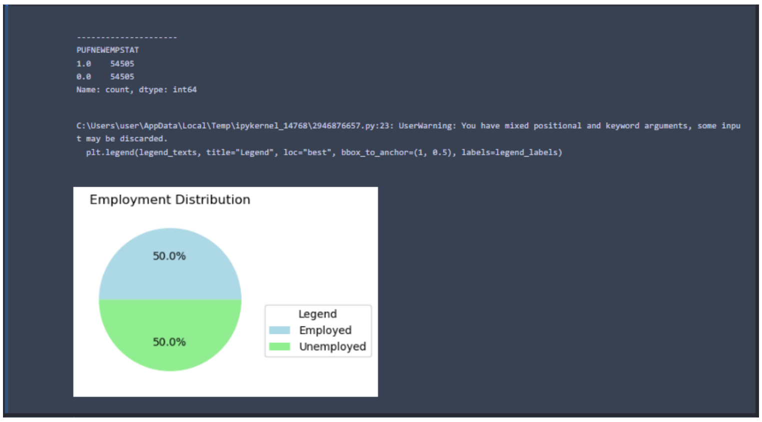


```
--------------------
PUFNEWEMPSTAT
1.0    54505
0.0     4952
Name: count, dtype: int64

C:\Users\user\AppData\Local\Temp\ipykernel_14768\405694498.py:18: UserWarning: You have mixed positional and keyword arguments, some input
may be discarded.
  plt.legend(legend_texts, title="Legend", loc="best", bbox_to_anchor=(1, 0.5), labels=legend_labels)
```

Since during exploratory analysis, the class distribution was discovered to have output binary imbalance. The researcher proceeded to preprocess the dataset through SMOTE to create synthetic data inputs and mitigate class imbalance issue and ensure that the analysis remains robust as shown below:



```
In [20]: # Create an instance of SMOTE
         from imblearn.over_sampling import SMOTE
         oversampler = SMOTE()

         # Resample data
         X_train_over, y_train_over = oversampler.fit_resample(X_train, y_train)

         print(dv)
         print(y_train_over.value_counts())

         # Create a pie chart
         plt.figure(figsize=(3, 3))
         class_counts = y_train_over.value_counts(normalize=True)*100
         colors = ['lightblue', 'lightgreen']

         plt.pie(class_counts, labels=['', ''], autopct='%1.1f%%', colors=colors)
         plt.title('Employment Distribution')

         # Create a custom legend
         legend_labels = ['Employed', 'Unemployed']
         legend_colors = colors
         legend_texts = [f'{label}: {class_counts[i]:.1%}' for i, label in enumerate(legend_labels)]
         plt.legend(legend_texts, title="Legend", loc="best", bbox_to_anchor=(1, 0.5), labels=legend_labels)

         plt.show()
```

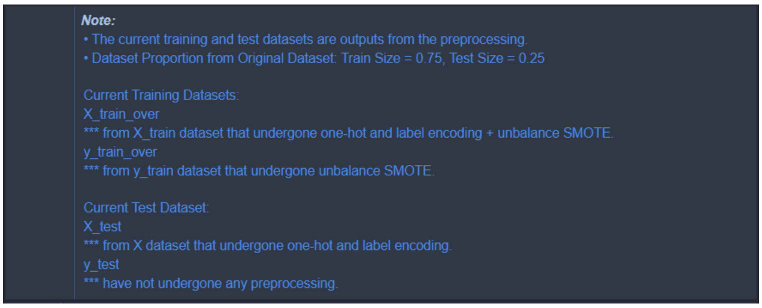


```
--------------------
PUFNEWEMPSTAT
1.0    54505
0.0    54505
Name: count, dtype: int64

C:\Users\user\AppData\Local\Temp\ipykernel_14768\2946876657.py:23: UserWarning: You have mixed positional and keyword arguments, some inpu
t may be discarded.
  plt.legend(legend_texts, title="Legend", loc="best", bbox_to_anchor=(1, 0.5), labels=legend_labels)
```

However, the researcher needs to exercise caution when using SMOTE, as it has the potential to introduce data noise and may lead to overfitting, making it challenging for models to generalize effectively to unseen data. Overfit models often perform poorly in real-world applications.

To address this concern, the researcher needs to conduct comparison between (1) SMOTE-processed 'X and y' training datasets and (2) 'X and y' training datasets before it was SMOTE-processed against the 'X and y' test dataset.

Preprocessing: Data Splitting (Train, Validation, Test)

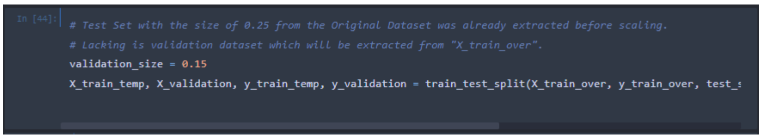


Note:
- The current training and test datasets are outputs from the preprocessing.
- Dataset Proportion from Original Dataset: Train Size = 0.75, Test Size = 0.25

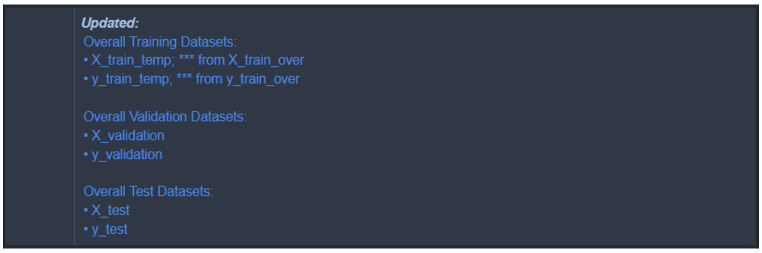
Current Training Datasets:
X_train_over
*** from X_train dataset that undergone one-hot and label encoding + unbalance SMOTE.
y_train_over
*** from y_train dataset that undergone unbalance SMOTE.

Current Test Dataset:
X_test
*** from X dataset that undergone one-hot and label encoding.
y_test
*** have not undergone any preprocessing.

Further data splitting was then conducted to have a validation dataset as shown below:



```
In [44]: # Test Set with the size of 0.25 from the Original Dataset was already extracted before scaling.
         # Lacking is validation dataset which will be extracted from "X_train_over".
         validation_size = 0.15
         X_train_temp, X_validation, y_train_temp, y_validation = train_test_split(X_train_over, y_train_over, test_s
```



Updated:

Overall Training Datasets:

- X_train_temp; *** from X_train_over
- y_train_temp; *** from y_train_over

Overall Validation Datasets:

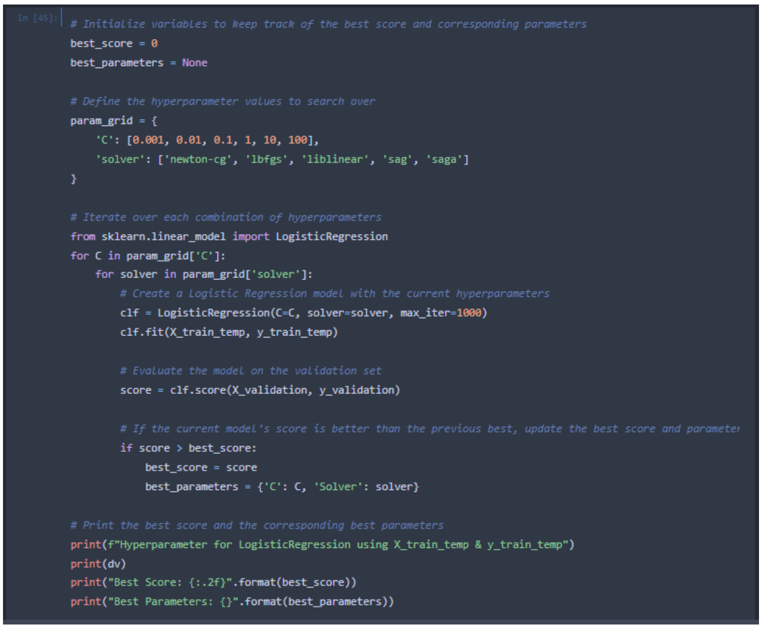
- X_validation
- y_validation

Overall Test Datasets:

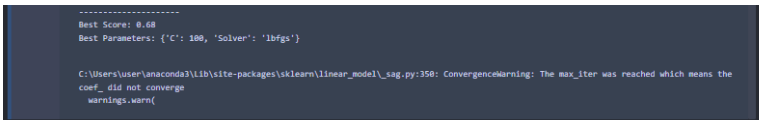
- X_test
- y_test

Hyperparameter Tuning was then carried out to find the best set of hyperparameters for a machine learning model, aiming to achieve optimal performance. This process involves a systematic search through a range of hyperparameter values and an evaluation of the model's performance for each combination as shown below:

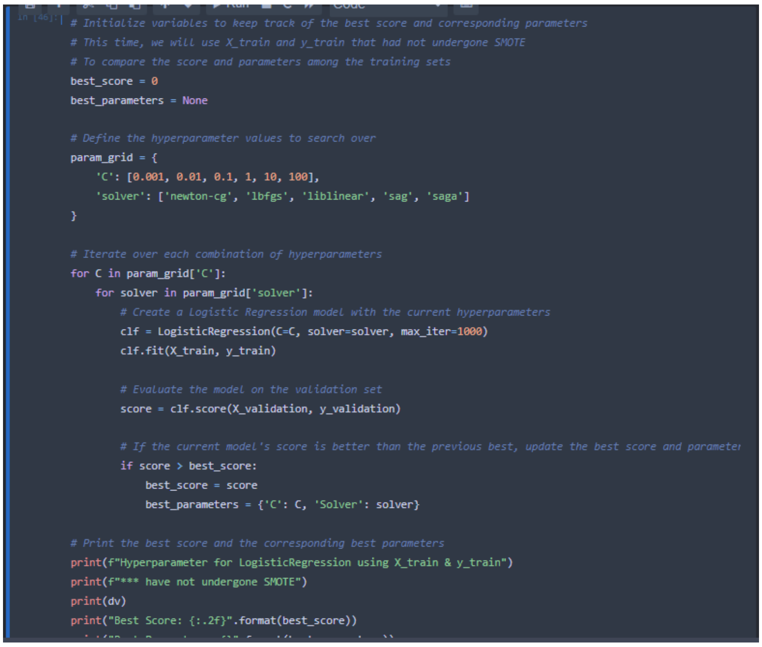
Hyperparameter Tuning



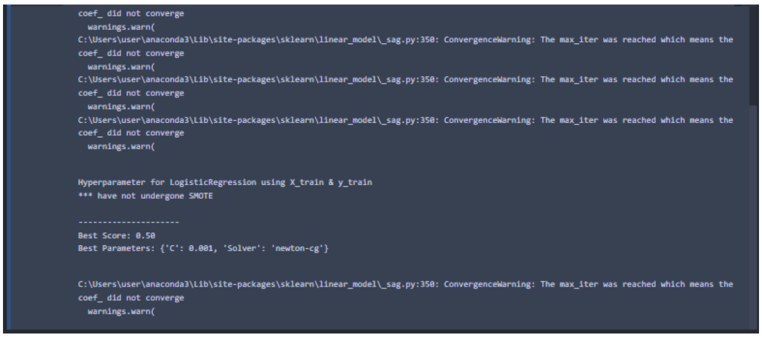
```
In [45]: # Initialize variables to keep track of the best score and corresponding parameters
         best_score = 0
         best_parameters = None

         # Define the hyperparameter values to search over
         param_grid = {
             'C': [0.001, 0.01, 0.1, 1, 10, 100],
             'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
         }

         # Iterate over each combination of hyperparameters
         from sklearn.linear_model import LogisticRegression
         for C in param_grid['C']:
             for solver in param_grid['solver']:
                 # Create a Logistic Regression model with the current hyperparameters
                 clf = LogisticRegression(C=C, solver=solver, max_iter=1000)
                 clf.fit(X_train_temp, y_train_temp)

                 # Evaluate the model on the validation set
                 score = clf.score(X_validation, y_validation)

                 # If the current model's score is better than the previous best, update the best score and parameter
                 if score > best_score:
                     best_score = score
                     best_parameters = {'C': C, 'Solver': solver}

         # Print the best score and the corresponding best parameters
         print(f"Hyperparameter for LogisticRegression using X_train_temp & y_train_temp")
         print(dv)
         print("Best Score: {:.2f}".format(best_score))
         print("Best Parameters: {}".format(best_parameters))
```



```
---------------------
Best Score: 0.68
Best Parameters: {'C': 100, 'Solver': 'lbfgs'}


C:\Users\user\anaconda3\Lib\site-packages\sklearn\linear_model\_sag.py:350: ConvergenceWarning: The max_iter was reached which means the
coef_ did not converge
  warnings.warn(
```



```python
# Initialize variables to keep track of the best score and corresponding parameters
# This time, we will use X_train and y_train that had not undergone SMOTE
# To compare the score and parameters among the training sets
best_score = 0
best_parameters = None

# Define the hyperparameter values to search over
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
}

# Iterate over each combination of hyperparameters
for C in param_grid['C']:
    for solver in param_grid['solver']:
        # Create a Logistic Regression model with the current hyperparameters
        clf = LogisticRegression(C=C, solver=solver, max_iter=1000)
        clf.fit(X_train, y_train)

        # Evaluate the model on the validation set
        score = clf.score(X_validation, y_validation)

        # If the current model's score is better than the previous best, update the best score and parameter
        if score > best_score:
            best_score = score
            best_parameters = {'C': C, 'Solver': solver}

# Print the best score and the corresponding best parameters
print(f"Hyperparameter for LogisticRegression using X_train & y_train")
print(f"*** have not undergone SMOTE")
print(dv)
print("Best Score: {:.2f}".format(best_score))
```

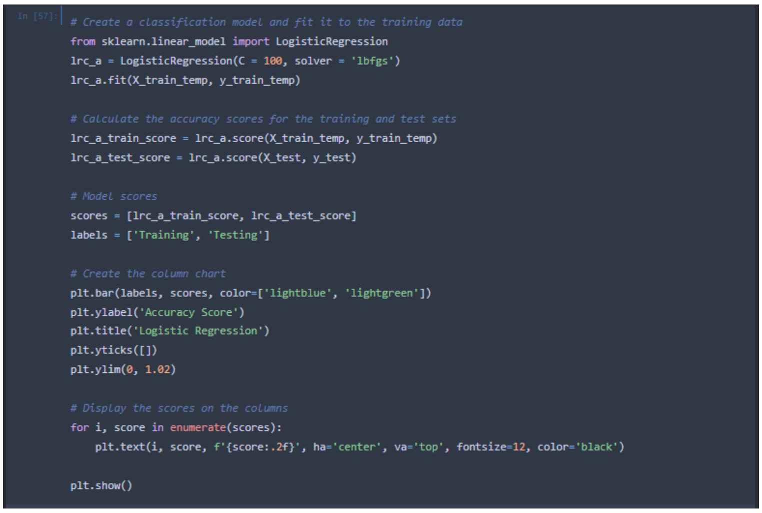


```
coef_ did not converge
  warnings.warn(
C:\Users\user\anaconda3\Lib\site-packages\sklearn\linear_model\_sag.py:350: ConvergenceWarning: The max_iter was reached which means the
coef_ did not converge
  warnings.warn(
C:\Users\user\anaconda3\Lib\site-packages\sklearn\linear_model\_sag.py:350: ConvergenceWarning: The max_iter was reached which means the
coef_ did not converge
  warnings.warn(
C:\Users\user\anaconda3\Lib\site-packages\sklearn\linear_model\_sag.py:350: ConvergenceWarning: The max_iter was reached which means the
coef_ did not converge
  warnings.warn(


Hyperparameter for LogisticRegression using X_train & y_train
*** have not undergone SMOTE

--------------------
Best Score: 0.50
Best Parameters: {'C': 0.001, 'Solver': 'newton-cg'}


C:\Users\user\anaconda3\Lib\site-packages\sklearn\linear_model\_sag.py:350: ConvergenceWarning: The max_iter was reached which means the
coef_ did not converge
  warnings.warn(
```

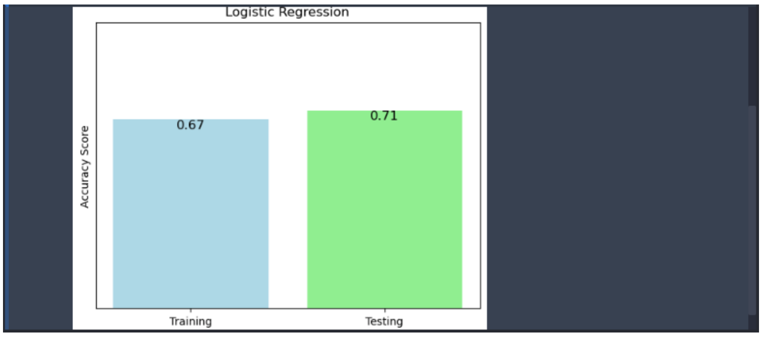
Two rounds of hyperparameter tuning were conducted to optimize the parameters for both the SMOTE-processed training dataset and the SMOTE-unprocessed training dataset. After hyperparameter tuning, the next steps involved model selection and training.

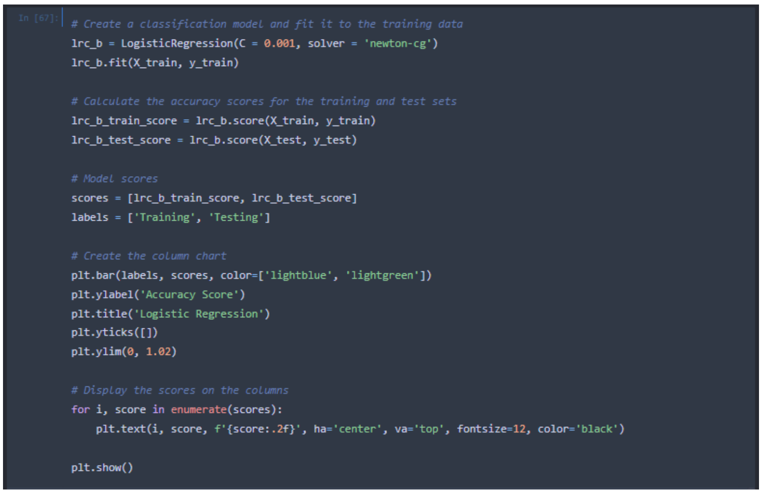
The researcher opted to use Logistic Regression for the model, with specific hyperparameters. For Logistic Regression A, which used the SMOTE-processed training dataset, the chosen hyperparameters were {“C” = 100, “solver” = "lbfgs"}. On the other hand, Logistic Regression B, utilizing the SMOTE-unprocessed training dataset, was configured with hyperparameters {“C” = 0.001, “solver” = "newton-cg"}. These hyperparameter choices were based on the results from the tuning process, as they achieved the best scores of 0.68 and 0.50, respectively.

Model Selection and Training

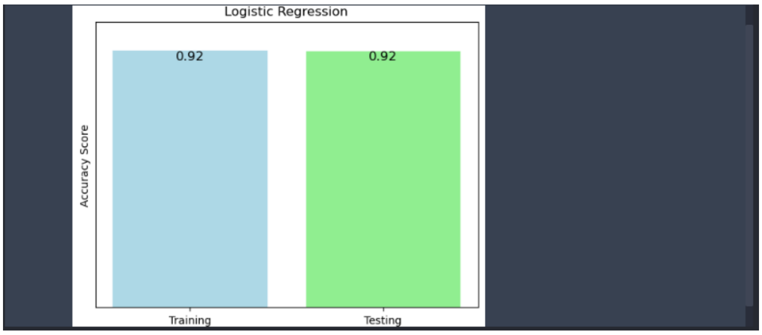


```
In [57]: # Create a classification model and fit it to the training data
         from sklearn.linear_model import LogisticRegression
         lrc_a = LogisticRegression(C = 100, solver = 'lbfgs')
         lrc_a.fit(X_train_temp, y_train_temp)

         # Calculate the accuracy scores for the training and test sets
         lrc_a_train_score = lrc_a.score(X_train_temp, y_train_temp)
         lrc_a_test_score = lrc_a.score(X_test, y_test)

         # Model scores
         scores = [lrc_a_train_score, lrc_a_test_score]
         labels = ['Training', 'Testing']

         # Create the column chart
         plt.bar(labels, scores, color=['lightblue', 'lightgreen'])
         plt.ylabel('Accuracy Score')
         plt.title('Logistic Regression')
         plt.yticks([])
         plt.ylim(0, 1.02)

         # Display the scores on the columns
         for i, score in enumerate(scores):
             plt.text(i, score, f'{score:.2f}', ha='center', va='top', fontsize=12, color='black')

         plt.show()
```





```
In [67]: # Create a classification model and fit it to the training data
         lrc_b = LogisticRegression(C = 0.001, solver = 'newton-cg')
         lrc_b.fit(X_train, y_train)

         # Calculate the accuracy scores for the training and test sets
         lrc_b_train_score = lrc_b.score(X_train, y_train)
         lrc_b_test_score = lrc_b.score(X_test, y_test)

         # Model scores
         scores = [lrc_b_train_score, lrc_b_test_score]
         labels = ['Training', 'Testing']

         # Create the column chart
         plt.bar(labels, scores, color=['lightblue', 'lightgreen'])
         plt.ylabel('Accuracy Score')
         plt.title('Logistic Regression')
         plt.yticks([])
         plt.ylim(0, 1.02)

         # Display the scores on the columns
         for i, score in enumerate(scores):
             plt.text(i, score, f'{score:.2f}', ha='center', va='top', fontsize=12, color='black')

         plt.show()
```



Surprisingly, the model fitted using the SMOTE-unprocessed training dataset performed remarkably well, achieving accuracy scores of 0.92 for both training and testing. In contrast, the model fitted using the SMOTE-processed training dataset exhibited lower performance, with accuracy scores of 0.68 for training and 0.71 for testing, respectively.

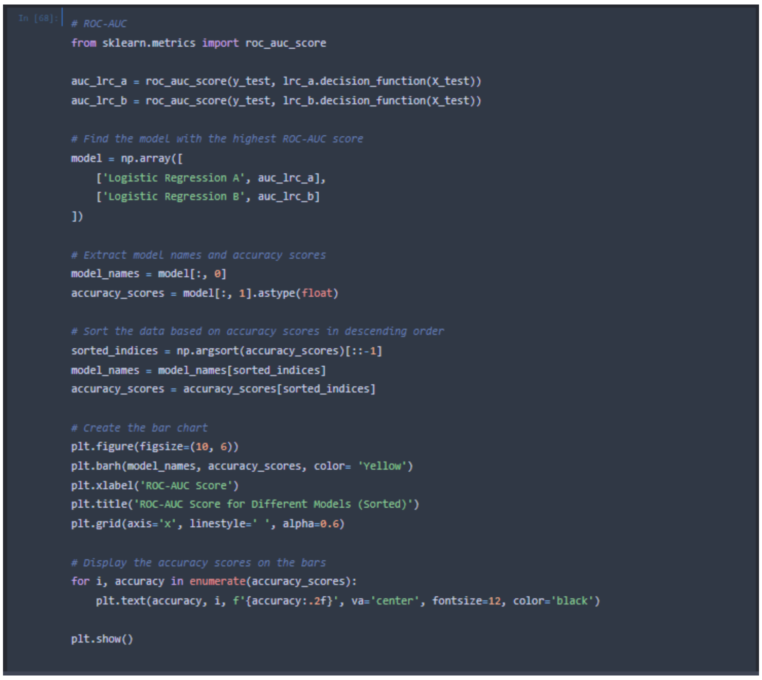
This striking contrast highlights the impact of applying SMOTE in our machine learning process. It suggests that while SMOTE can be a valuable tool for addressing class imbalance, its application doesn't always guarantee improved model performance. In some cases, as observed here, the original unbalanced dataset may provide more reliable results. This outcome emphasizes the importance of carefully considering the specific dataset and its characteristics when deciding whether to employ SMOTE in a machine learning project.

After fitting the models and obtaining the accuracy scores, the next step involved comprehensive Model Evaluation. This evaluation process aimed to provide a more holistic understanding of how well the models were performing.

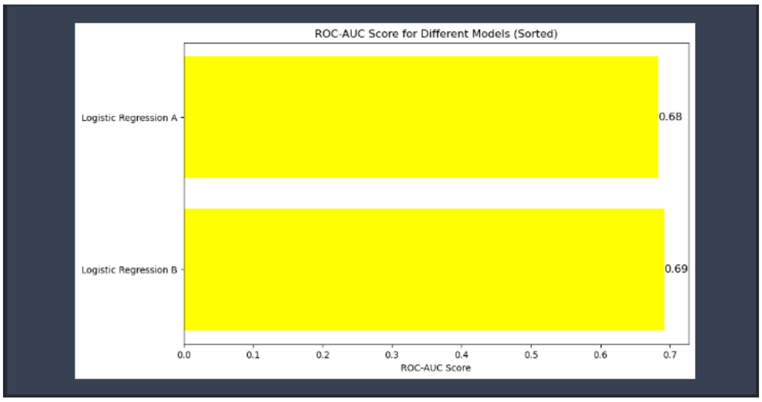
One of the key metrics used for assessment was the Receiver Operating Characteristic Area Under the Curve (ROC-AUC). The ROC-AUC is a measure of a model's ability to distinguish between the positive and negative classes. It provides valuable insights into the model's overall discriminatory power, making it a crucial metric for binary classification tasks.

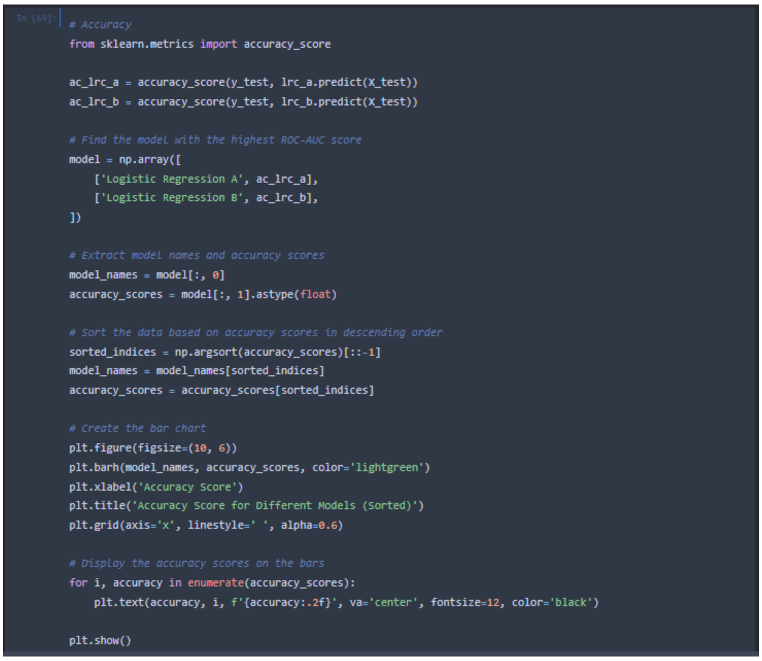
Additionally, accuracy and F-1 score were examined. Accuracy, as previously mentioned, indicates the proportion of correctly classified instances, serving as a general indicator of a model's overall correctness. On the other hand, the F-1 score combines precision and recall, offering a more balanced evaluation, especially when dealing with imbalanced datasets.

Model Evaluation

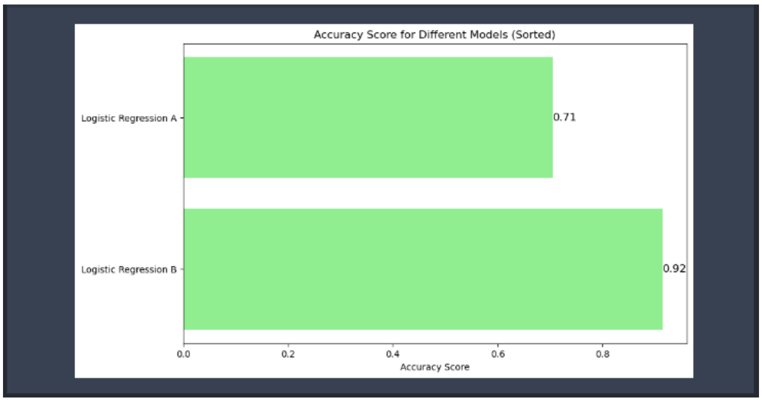


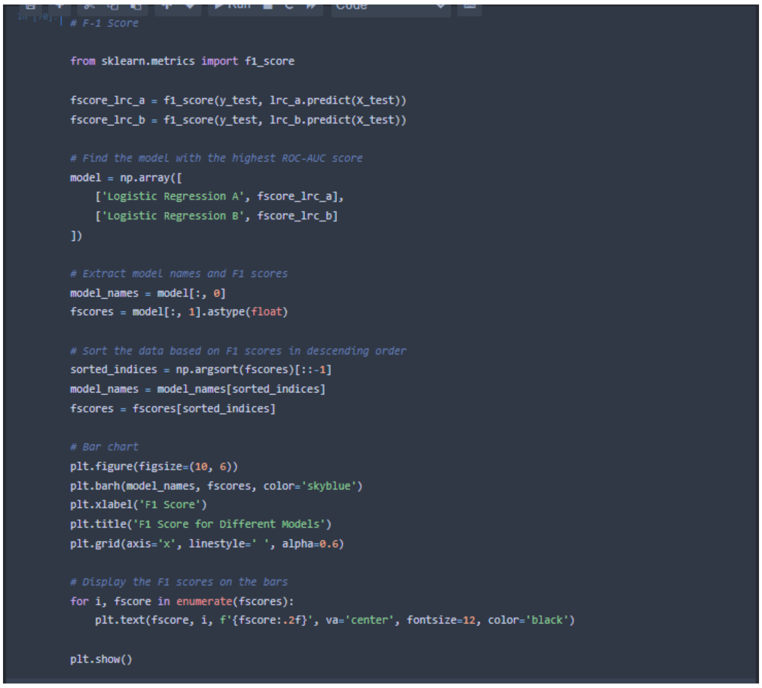
```
In [68]: # ROC-AUC
         from sklearn.metrics import roc_auc_score

         auc_lrc_a = roc_auc_score(y_test, lrc_a.decision_function(X_test))
         auc_lrc_b = roc_auc_score(y_test, lrc_b.decision_function(X_test))

         # Find the model with the highest ROC-AUC score
         model = np.array([
             ['Logistic Regression A', auc_lrc_a],
             ['Logistic Regression B', auc_lrc_b]
         ])

         # Extract model names and accuracy scores
         model_names = model[:, 0]
         accuracy_scores = model[:, 1].astype(float)

         # Sort the data based on accuracy scores in descending order
         sorted_indices = np.argsort(accuracy_scores)[::-1]
         model_names = model_names[sorted_indices]
         accuracy_scores = accuracy_scores[sorted_indices]

         # Create the bar chart
         plt.figure(figsize=(10, 6))
         plt.barh(model_names, accuracy_scores, color= 'Yellow')
         plt.xlabel('ROC-AUC Score')
         plt.title('ROC-AUC Score for Different Models (Sorted)')
         plt.grid(axis='x', linestyle=' ', alpha=0.6)

         # Display the accuracy scores on the bars
         for i, accuracy in enumerate(accuracy_scores):
             plt.text(accuracy, i, f'{accuracy:.2f}', va='center', fontsize=12, color='black')

         plt.show()
```



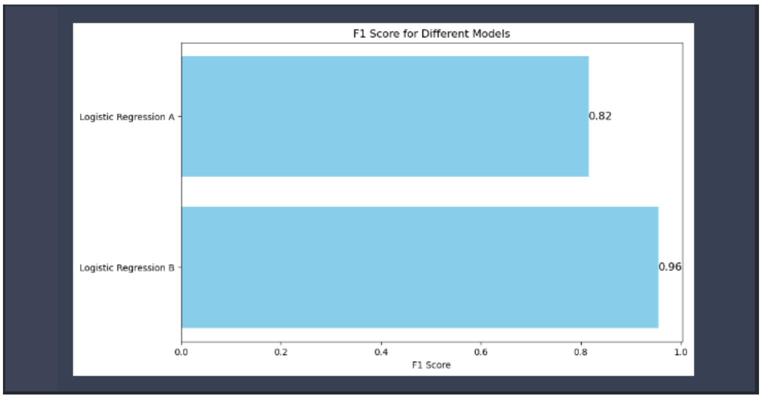


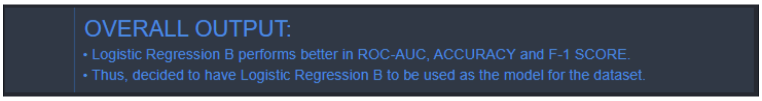
```
# Accuracy
from sklearn.metrics import accuracy_score

ac_lrc_a = accuracy_score(y_test, lrc_a.predict(X_test))
ac_lrc_b = accuracy_score(y_test, lrc_b.predict(X_test))

# Find the model with the highest ROC-AUC score
model = np.array([
    ['Logistic Regression A', ac_lrc_a],
    ['Logistic Regression B', ac_lrc_b],
])

# Extract model names and accuracy scores
model_names = model[:, 0]
accuracy_scores = model[:, 1].astype(float)

# Sort the data based on accuracy scores in descending order
sorted_indices = np.argsort(accuracy_scores)[::-1]
model_names = model_names[sorted_indices]
accuracy_scores = accuracy_scores[sorted_indices]

# Create the bar chart
plt.figure(figsize=(10, 6))
plt.barh(model_names, accuracy_scores, color='lightgreen')
plt.xlabel('Accuracy Score')
plt.title('Accuracy Score for Different Models (Sorted)')
plt.grid(axis='x', linestyle=' ', alpha=0.6)

# Display the accuracy scores on the bars
for i, accuracy in enumerate(accuracy_scores):
    plt.text(accuracy, i, f'{accuracy:.2f}', va='center', fontsize=12, color='black')

plt.show()
```





```
# F-1 Score

from sklearn.metrics import f1_score

fscore_lrc_a = f1_score(y_test, lrc_a.predict(X_test))
fscore_lrc_b = f1_score(y_test, lrc_b.predict(X_test))

# Find the model with the highest ROC-AUC score
model = np.array([
    ['Logistic Regression A', fscore_lrc_a],
    ['Logistic Regression B', fscore_lrc_b]
])

# Extract model names and F1 scores
model_names = model[:, 0]
fscores = model[:, 1].astype(float)

# Sort the data based on F1 scores in descending order
sorted_indices = np.argsort(fscores)[::-1]
model_names = model_names[sorted_indices]
fscores = fscores[sorted_indices]

# Bar chart
plt.figure(figsize=(10, 6))
plt.barh(model_names, fscores, color='skyblue')
plt.xlabel('F1 Score')
plt.title('F1 Score for Different Models')
plt.grid(axis='x', linestyle=' ', alpha=0.6)

# Display the F1 scores on the bars
for i, fscore in enumerate(fscores):
    plt.text(fscore, i, f'{fscore:.2f}', va='center', fontsize=12, color='black')

plt.show()
```




OVERALL OUTPUT:

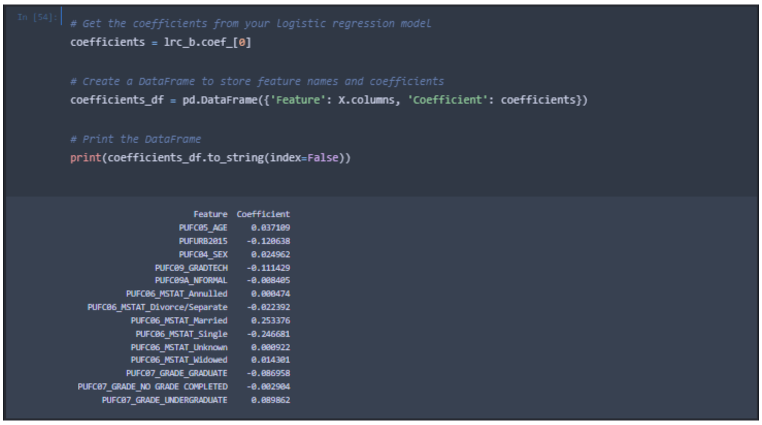
- Logistic Regression B performs better in ROC-AUC, ACCURACY and F-1 SCORE.
- Thus, decided to have Logistic Regression B to be used as the model for the dataset.

Overall, the comparison between Logistic Regression B, fitted with SMOTE-unprocessed training datasets, and Logistic Regression A, fitted with SMOTE-processed training datasets, reveals that Logistic Regression B outperformed Logistic Regression A in multiple key metrics.

Logistic Regression B achieved a higher ROC-AUC score of 0.69 compared to 0.68 for Logistic Regression A, indicating a better ability to distinguish between positive and negative classes. Moreover, Logistic Regression B demonstrated superior accuracy, scoring 0.92 compared to 0.71 for Logistic Regression A. The F1 Score, which balances precision and recall, was also substantially higher for Logistic Regression B, measuring 0.96 compared to 0.82 for Logistic Regression A.

As a result of these superior performance metrics, the researcher made the decision to deploy Logistic Regression B as the chosen model for predicting the likelihood of a person being employed. This selection was based on Logistic Regression B's stronger overall performance, emphasizing the significance of ROC-AUC score, accuracy, and F1 Score in the decision-making process.

Model Deployment



```
In [54]: # Get the coefficients from your logistic regression model
         coefficients = lrc_b.coef_[0]

         # Create a DataFrame to store feature names and coefficients
         coefficients_df = pd.DataFrame({'Feature': X.columns, 'Coefficient': coefficients})

         # Print the DataFrame
         print(coefficients_df.to_string(index=False))
```

```
                        Feature  Coefficient
                     PUFC05_AGE     0.037109
                     PUFURB2015    -0.120638
                     PUFC04_SEX     0.024962
                PUFC09_GRADTECH    -0.111429
                PUFC09A_NFORMAL    -0.008405
          PUFC06_MSTAT_Annulled     0.000474
  PUFC06_MSTAT_Divorce/Separate    -0.022392
           PUFC06_MSTAT_Married     0.253376
            PUFC06_MSTAT_Single    -0.246681
           PUFC06_MSTAT_Unknown     0.000922
           PUFC06_MSTAT_Widowed     0.014301
          PUFC07_GRADE_GRADUATE    -0.086958
PUFC07_GRADE_NO GRADE COMPLETED    -0.002904
     PUFC07_GRADE_UNDERGRADUATE     0.089862
```

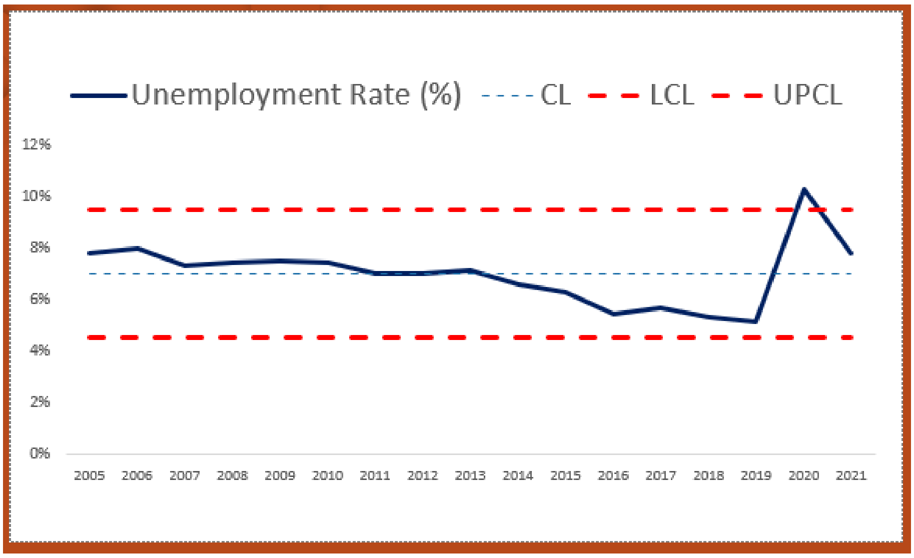
The coefficients presented above hold valuable information about the likelihood of certain outcomes based on the features listed on their left side. In logistic regression, these coefficients help us understand the impact of each feature on the probability of a particular event occurring.

For example, a positive coefficient indicates that an increase in the corresponding feature will raise the likelihood of the event happening, while a negative coefficient suggests that an increase in the feature will lower the likelihood. The magnitude of the coefficient also matters; a larger coefficient signifies a stronger influence on the outcome.

RESULTS & DISCUSSION

Portraying the Unemployment Landscape: (Descriptive Analytics)

- ***What is the current unemployment rate in the Philippines, and how has it changed over time?***



The unemployment rate declined steadily from 2005 to 2019 with small fluctuations between lower and upper control limit until reaching a low of 5.11% in 2019. However, in 2020, the unemployment rate (10.26%) went above the upper limit. This indicates an issue, and this means a thorough investigation is needed to determine the cause of this high rate.

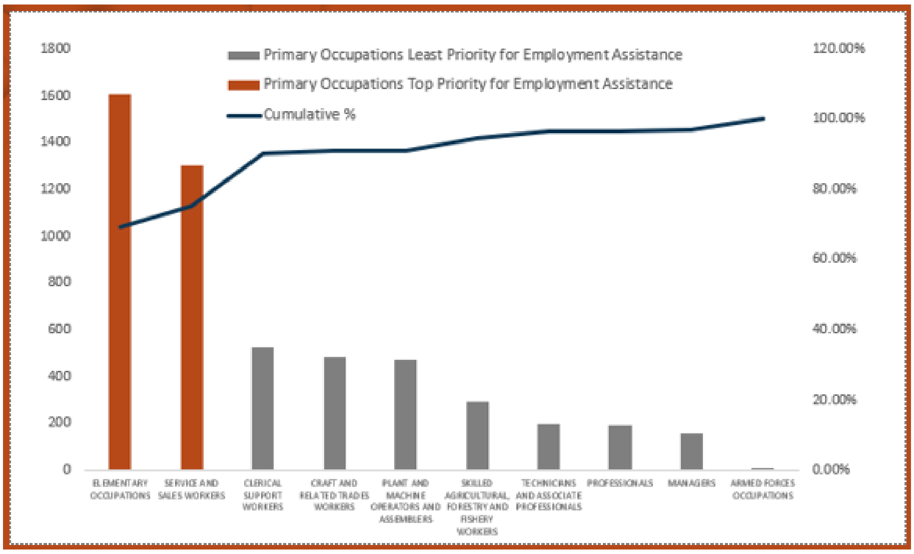
Upon further investigation, we commonly know that at 2020, the world faced an unprecedented challenge in the form of the COVID-19 pandemic. The pandemic had far-reaching effects on the global and local economy, leading to business closures, job losses, and economic uncertainty. This event likely played a significant role in the sharp increase in the unemployment rate in that year.

Also, this means that the data suggests that prior to 2020, the economy had been relatively stable, with the unemployment rate staying within a manageable range. However, the sudden spike in 2020 highlights the importance of understanding and preparing for unforeseen events that can disrupt economic stability.

Overall Findings:

- Prior to 2020, indications of economic stability are observed due to a steady decline in unemployment rate (2005 to 2019) with fluctuations between lower and upper control limit.
- Sharp increase in unemployment rate in 2020 due to COVID-19 pandemic that went above the upper limit.
- Importance of understanding and preparing for unforeseen events that can disrupt economic stability.

- ***What are the top primary occupations in need of assistance when people are searching for employment?***



It's evident that the primary contributors to the overall unemployment issue are the occupations of "Elementary Occupations" and "Service and Sales Workers." Together, these two categories account for a significant share of the unemployment rate, making up as much as 75% of the total unemployment among all the listed occupations.

This means that a large proportion of people without jobs are found in these two occupational groups. In simpler terms, if we were to group all the unemployed individuals by their primary occupation, we'll find that the majority fall into either "Elementary Occupations" or "Service and Sales Workers."

In contrast, the remaining occupations listed on the chart have notably smaller contributions to the overall unemployment rate. While they still play a part in the unemployment issue, their impact is relatively minor compared to the significant presence of unemployed individuals in the "Elementary Occupations" and "Service and Sales Workers" categories.

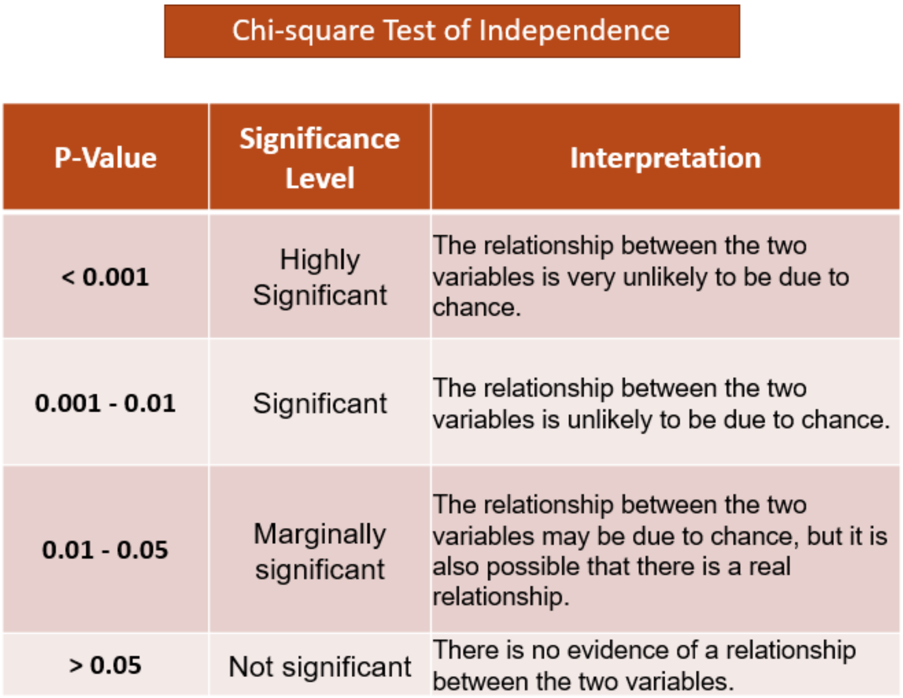
Overall Findings:

- “Elementary Occupations” and “Service and Sales Workers” account for 75% of total unemployment.
- This means that a large proportion of unemployed individuals are found in the above two occupational groups.
- Other occupations have notably smaller contributions to the overall unemployment rate.

Diagnosing the Unemployment Ailments: (Diagnostic Analytics)

- ***Why is there a variation in employment status among different demographic in the Philippines, and are there specific factors contributing to this discrepancy?***

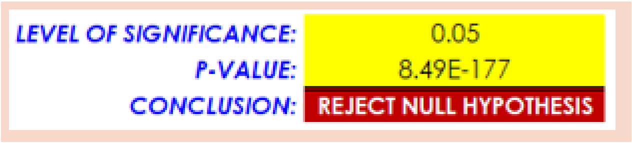
Through the use of a Chi-square test, we can identify crucial factors that have a substantial impact on employment status in the Philippines. We assess their significance through the p-values generated in our Chi-square test of independence as shown below:



Chi-square Test of Independence

P-Value	Significance Level	Interpretation
< 0.001	Highly Significant	The relationship between the two variables is very unlikely to be due to chance.
0.001 - 0.01	Significant	The relationship between the two variables is unlikely to be due to chance.
0.01 - 0.05	Marginally significant	The relationship between the two variables may be due to chance, but it is also possible that there is a real relationship.
> 0.05	Not significant	There is no evidence of a relationship between the two variables.

Region Demographic Profile

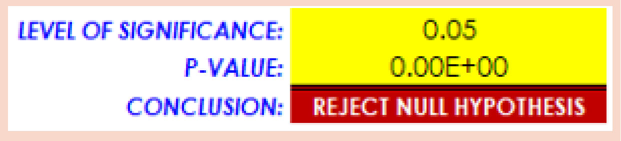


LEVEL OF SIGNIFICANCE:	0.05
P-VALUE:	8.49E-177
CONCLUSION:	REJECT NULL HYPOTHESIS

"Reject null hypothesis" means that the statistical analysis has shown that there is a significant relationship between the region profile and their employment status due to the significantly small “8.49E-177” p-value.

In simpler terms, it confirms that where you live has a meaningful impact on whether you have a job or not. Different regions in the Philippines may have varying levels of economic activity, industries, and job opportunities, leading to distinct employment outcomes.

Highest Grade Completed

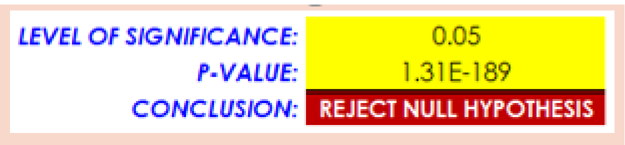


LEVEL OF SIGNIFICANCE:	0.05
P-VALUE:	0.00E+00
CONCLUSION:	REJECT NULL HYPOTHESIS

"Reject null hypothesis" means that the statistical analysis has revealed a significant relationship between the highest grade completed and employment status due to the significantly small “0.00E+00” p-value.

In simpler terms, it confirms that the level of education you attain significantly influences whether you have a job or not. Those with higher levels of education may have better skills and qualifications, making them more employable.

Technical Vocational Graduate



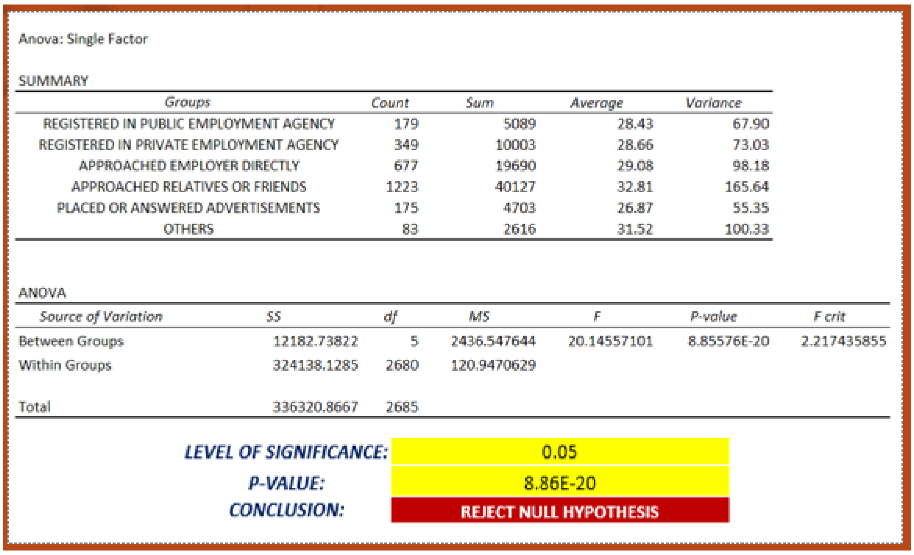
LEVEL OF SIGNIFICANCE:	0.05
P-VALUE:	1.31E-189
CONCLUSION:	REJECT NULL HYPOTHESIS

"Reject null hypothesis" means that the statistical analysis has shown a significant relationship between completing technical or vocational courses and employment status due to the significantly small "1.31E-189" p-value.

In simpler terms, it confirms that completing these courses significantly influences whether you have a job or not. This suggests that individuals who have completed technical or vocational courses have distinct employment patterns.

- ***Why do different age groups have their unique ways of job hunting, and can we figure out what's causing these differences?***

Much like the Chi-square test helps uncover vital factors influencing employment status in the Philippines, we use a statistical tool called One Way ANOVA to explore the distinct job-hunting behaviors across different age groups.



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
REGISTERED IN PUBLIC EMPLOYMENT AGENCY	179	5089	28.43	67.90
REGISTERED IN PRIVATE EMPLOYMENT AGENCY	349	10003	28.66	73.03
APPROACHED EMPLOYER DIRECTLY	677	19690	29.08	98.18
APPROACHED RELATIVES OR FRIENDS	1223	40127	32.81	165.64
PLACED OR ANSWERED ADVERTISEMENTS	175	4703	26.87	55.35
OTHERS	83	2616	31.52	100.33

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	12182.73822	5	2436.547644	20.14557101	8.85576E-20	2.217435855
Within Groups	324138.1285	2680	120.9470629			
Total	336320.8667	2685				

LEVEL OF SIGNIFICANCE:	0.05
P-VALUE:	8.86E-20
CONCLUSION:	**REJECT NULL HYPOTHESIS**

The ANOVA analysis results are quite remarkable. In this ANOVA analysis, we're checking how different groups approach job searching. The groups include those who register in public or private employment agencies, approach employers directly, seek help from relatives or friends, respond to advertisements, or use other methods.

Looking at the results, we see two main parts: "Between Groups" and "Within Groups." "Between Groups" shows differences in job search methods among these categories, while "Within Groups" considers differences within each group.

The low p-value (8.86E-20) and significant F-value (20.15) indicate that the diverse job search methods among these groups are not random; there's a real difference. By rejecting the null hypothesis, we confirm that these groups indeed have unique approaches to job searching.

In simpler terms, this analysis tells us that people in these categories don't search for jobs in the same way, and this understanding can help create targeted strategies to assist them better.

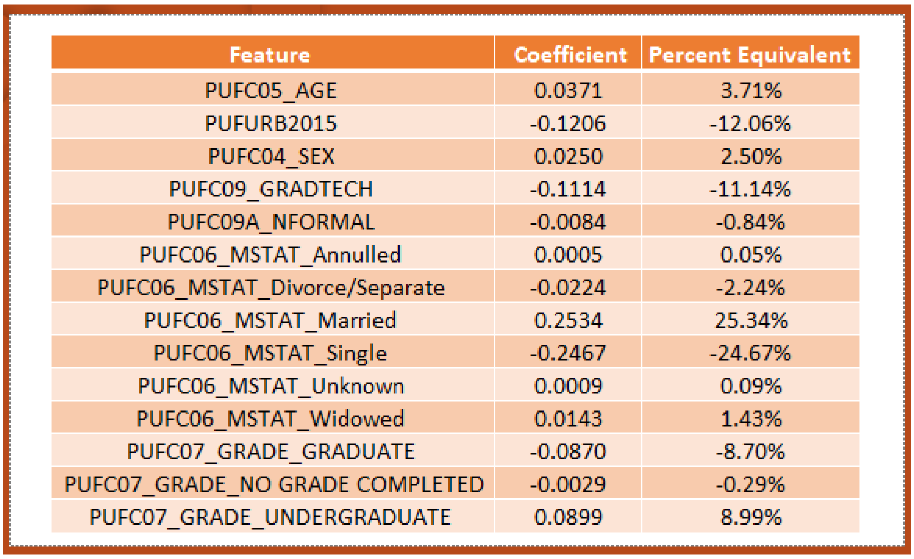
Overall Findings:

- ANOVA reveals significant and non-random differences in job search methods across different age groups. (p-value = 8.86E-20)
- People of different age groups exhibit unique and distinct job-hunting behaviors.
- Age has a significant influence on how people search for jobs.
- "Between Groups" and "Within Groups" distinctions showcase differences in job search methods among categories and within each group, respectively.

Predicting the Future of Unemployment: (Predictive Analytics)

- ***Can we predict the likelihood of an individual in the Philippines being employed or unemployed?***

We used logistic regression analysis to predict the likelihood of a binary outcome – in this case, whether a person will find employment or not. The output provides coefficients and estimates for each predictor variable, which helps us identify the key factors that impact employment status as shown below:

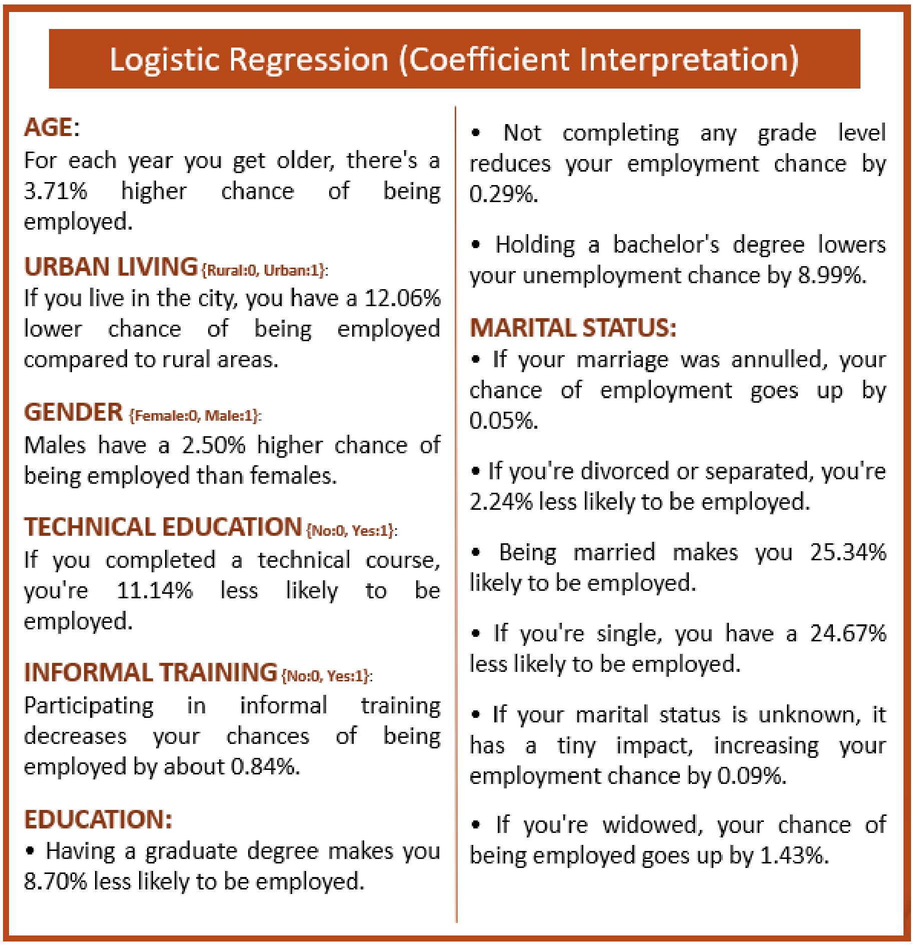


Feature	Coefficient	Percent Equivalent
PUFC05_AGE	0.0371	3.71%
PUFURB2015	-0.1206	-12.06%
PUFC04_SEX	0.0250	2.50%
PUFC09_GRADTECH	-0.1114	-11.14%
PUFC09A_NFORMAL	-0.0084	-0.84%
PUFC06_MSTAT_Annulled	0.0005	0.05%
PUFC06_MSTAT_Divorce/Separate	-0.0224	-2.24%
PUFC06_MSTAT_Married	0.2534	25.34%
PUFC06_MSTAT_Single	-0.2467	-24.67%
PUFC06_MSTAT_Unknown	0.0009	0.09%
PUFC06_MSTAT_Widowed	0.0143	1.43%
PUFC07_GRADE_GRADUATE	-0.0870	-8.70%
PUFC07_GRADE_NO GRADE COMPLETED	-0.0029	-0.29%
PUFC07_GRADE_UNDERGRADUATE	0.0899	8.99%

Discussion:

The above features and coefficients represent the various factors that can affect the likelihood of unemployment. However, the findings suggesting that having a technical course, informal training, graduate degree and living in the city make you less likely to be employed may raise questions.

This means, it's crucial to consider the dataset and the variables in play. Further analysis and investigation are needed for a deeper understanding of these observations.



Logistic Regression (Coefficient Interpretation)

AGE:
For each year you get older, there's a 3.71% higher chance of being employed.

URBAN LIVING {Rural:0, Urban:1}:
If you live in the city, you have a 12.06% lower chance of being employed compared to rural areas.

GENDER {Female:0, Male:1}:
Males have a 2.50% higher chance of being employed than females.

TECHNICAL EDUCATION {No:0, Yes:1}:
If you completed a technical course, you're 11.14% less likely to be employed.

INFORMAL TRAINING {No:0, Yes:1}:
Participating in informal training decreases your chances of being employed by about 0.84%.

EDUCATION:

- Having a graduate degree makes you 8.70% less likely to be employed.
- Not completing any grade level reduces your employment chance by 0.29%.
- Holding a bachelor's degree lowers your unemployment chance by 8.99%.

MARITAL STATUS:

- If your marriage was annulled, your chance of employment goes up by 0.05%.
- If you're divorced or separated, you're 2.24% less likely to be employed.
- Being married makes you 25.34% likely to be employed.
- If you're single, you have a 24.67% less likely to be employed.
- If your marital status is unknown, it has a tiny impact, increasing your employment chance by 0.09%.
- If you're widowed, your chance of being employed goes up by 1.43%.

CONCLUSION & RECOMMENDATION

UNEMPLOYMENT RATE

Conclusion:

ECONOMIC STABILITY PRE-2020: The control chart analysis visually illustrates fluctuations in unemployment rates. The variations observed before year 2020 are likely attributed to common causes inherent in the process, such as economic cycles or seasonal employment changes.

COVID-19 IMPACT: The sharp increase in the 2020 unemployment rate was caused by the impact of COVID-19 on the global and local economy. This led to businesses closing down, people losing their jobs, and economic uncertainty. This unexpected event greatly disturbed the previously steady economic conditions.

Recommendation:

Given the abrupt spike in the 2020 unemployment rate beyond the established limits, it is crucial to implement proactive measures for economic resilience. This includes developing contingency plans and policies that address the impact of unforeseen events, such as global crises or pandemics.

Additionally, continuous monitoring and analysis of economic indicators can aid in early detection of potential issues, allowing for timely intervention and mitigation strategies. By enhancing preparedness and adaptability, the economy can better withstand external shocks and maintain stability in the face of uncertainties.

UNEMPLOYED OCCUPATIONS

Conclusion:

TARGETED INTERVENTIONS: After analyzing the Pareto chart, it's clear that the key players in the unemployment problem are the jobs in "Elementary Occupations" and "Service and Sales Workers." Together, they make up a whopping 75% of all the unemployment cases.

In simple terms, most of the people without jobs are either in "Elementary Occupations" or "Service and Sales Workers." Other jobs on the list contribute less to the overall

MINOR CONTRIBUTIONS: Even though other listed occupations are small in value, they still have some relative impact overall.

Recommendation:

Based on the Pareto Chart Analysis, the major culprits behind unemployment are "Elementary Occupations" and "Service and Sales Workers," making up a whopping 75% of the total unemployment rate. To address this, we can tailor our plans for each job type and focus on the most significant contributors to effectively combat unemployment.

Also, understanding the unique challenges each job faces is crucial which allows us to create tailored rules and plans that cater to their specific needs. These approaches aim to build a well-rounded job market that can handle challenges more effectively, ultimately reducing overall unemployment.

FACTORS IMPACTING EMPLOYMENT

Conclusion:

CRUCIAL FACTORS REVEALED BY CHI-SQUARE TEST: Significance is gauged through p-values, where smaller values indicate greater influence on job prospects. This means that the influential factors act as keys shaping job opportunities and vary based on our background related to them; such as follows:

- Different regions have varying economic activity, industries, and job opportunities, leading to diverse employment outcomes.
- Higher education levels may result in better skills and qualifications, enhancing employability.
- Individuals with technical or vocational training exhibit distinct employment patterns.

Recommendation:

For better job opportunities, policymakers need to focus on improving underdeveloped regions by increasing economic activity, creating new industries, and generating jobs. On the other hand, educators also have a key role in this by ensuring students receive quality education and training that equips them with the skills needed for the job market.

In regards with technical and vocational education programs, investment is crucial as completing these courses significantly impacts employment status. This way, people can gain the skills and knowledge necessary for success in the job market.

AGE & JOB SEARCH TREND

Conclusion:

SIGNIFICANT FINDINGS: The ANOVA analysis yielded compelling results, with a remarkably small p-value (8.86E-20) suggesting a non-random disparity in how distinct age groups approach job searching, leading to the rejection of the "NULL" hypothesis.

Essentially, this confirms that different age groups employ unique and distinct job-hunting methods.

VALUABLE FOUNDATION: The understanding from ANOVA analysis provides a solid foundation for the development of targeted programs and services aimed at assisting job seekers across all age groups.

Recommendation:

As revealed by the ANOVA analysis, people in various categories has a unique job search behaviors and approach. This means, it's crucial to design targeted programs that cater to their specific needs such as:

- Younger job seekers might focus on online job platforms.
- Older job seekers could emphasize networking and traditional job search methods.

This tailored approach ensures that support programs resonate with the distinct job-hunting behaviors exhibited by individuals of different age groups, ultimately enhancing the effectiveness of assistance initiatives.

PREDICTING EMPLOYMENT STATUS

Conclusion:

SIGNIFICANT FINDINGS: The logistic regression analysis revealed that age, urban living, gender, technical education, informal training, and education level are significant factors that impact employment status. Marital status was also found to be a significant factor, with being married decreasing the likelihood of unemployment while being single increases it.

VALUABLE FOUNDATION: The analysis provides valuable insights into the factors that impact employment status and can be used to design targeted policies, programs, and initiatives to assist job seekers.

Recommendation:

Policymakers and stakeholders can develop targeted programs and services to help job seekers of different ages, genders, and educational backgrounds such as:

- Tailored job placement programs for different age groups can help individuals find suitable employment opportunities more effectively.

Also, employers can create more inclusive hiring practices such as gender-inclusive employment policies. Address the 2.50% higher chance of employment for males by promoting gender equality and supporting industries where gender disparities exist.

Meanwhile, in-depth analysis and investigation should be conducted to validate and refine the observations from the dataset since having a technical course, informal training, graduate degree and living in the city make you less likely to be employed may raise questions.

REFERENCES AND BIBLIOGRAPHY

Philippines, Key Indicators 2023 (XLSX) | ADB Data Library | Asian Development Bank. (n.d.). https://data.adb.org/media/11466

Labor Force Survey | Philippine Statistics Authority | Republic of the Philippines. (2023, November 8). https://psa.gov.ph/statistics/labor-force-survey

Salvosa, F. (2015, September 2). Philippines struggles with unemployment despite economic growth. *CNBC*. https://www.cnbc.com/2015/09/01/unemployment-in-philippines-an-issue-despite-rapid-economic-growth.html

Ligot, D. V., Melendres, R. L., Tayco, F. C., Vizmonte, E. J., Toledo, M., Gerlock-Barretto, A., Martínez, J. A., Bernardo, G., Sindol-Ritualo, M., Néri, C., Bungcaras, J., & Pelayo, S. (2022b). Philippines Data Analytics Sector Labor Market Intelligence Report. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4027384

Smaldone, F., Ippolito, A., Lagger, J., & Pellicano, M. (2022). Employability skills: Profiling data scientists in the digital labour market. *European Management Journal*, *40*(5), 671–684. https://doi.org/10.1016/j.emj.2022.05.005

Why should we integrate income and employment support? A conceptual and empirical investigation. (n.d.). https://www.ilo.org/static/english/intserv/working-papers/wp072/index.html

Brooks, R., & Author_Id, N. (2002). Why is Unemployment High in the Philippines? *IMF Working Paper*, *02*(23), 1. https://doi.org/10.5089/9781451844054.001