Team 05

BUA4004
Teamwork Assessment
Data Analysis Report

Team Members

1.) Rajlakshmi Varma (21999137) – Team Leader
2.) Sathavara Suchi (22237665)
3.) Ruchika Parekh (22173623)
4.) Sabin Ghimire (22386922)

Introduction: For any analysis cycle getting accurate data is first and most important thing. To achieve this, we need to perform several actions on current dataset like preparing, cleaning & manipulating. After that model building and analysis comes in the picture. Last and important cycle is to communicate that finding for better and useful suggestion.

Task 1

In this task we have filtered the dataset to include only children aged 6 to 17 who are identified as the children of the household head. We used **IF(AND(R6>=6, R6<=17, T12="Child"), 1, 0)** in a new column, each row will now have a **1** (if it meets the criteria) or **0** (if it does not). The final dataset will contain only the relevant children, reducing errors and improving the accuracy of your analysis.

Task 1.2:

We need to ensure that our analysis only focus on children appropriate for schooling age thereby increasing accuracy and relatability of our results. By doing so we can focus more on other factors that affects children's schooling.

We use below formula and save the results into the column, IF(AND(R6=6, AO6="Too young"), 1,0) in **"Age_School_Filter"** . If the child is 6-year-old and classified as "Too young" it will return the value 1 and if not, then 0. After getting these values filter out row which has value 1 in "Age_School_Filter".

Task 2

For further analysis we are going to check if age and gender plays any sort of role for the children being out of school.

Task 2.1:

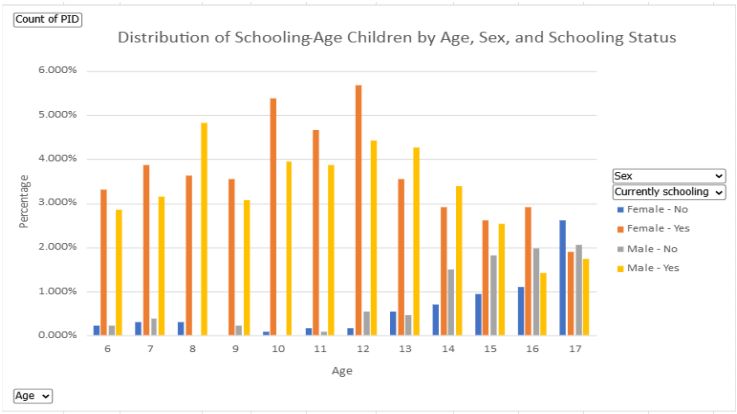
To calculate relative frequency between those factors we need to ensure that data is cleaned. In "Currently schooling" there are missing value to replace them by NO we used **replace all** function and made sure there is no blank cell in our dataset.

Pivot Table: We selected data and made pivot table. In that we added Age in row, Currently_schooling and Sex as column and Count of PID in values. After making table show the value as **% of grand total** from 'show value as' and round up the figure by **3 decimal** using format cells.

Count of PID	Column Labels						
	Female		Female Total	Male		Male Total	Grand Total
Row Labels	No	Yes		No	Yes		
6	0.237%	3.318%	3.555%	0.237%	2.844%	3.081%	6.635%
7	0.316%	3.870%	4.186%	0.395%	3.160%	3.555%	7.741%
8	0.316%	3.633%	3.949%	0.000%	4.818%	4.818%	8.768%
9	0.000%	3.555%	3.555%	0.237%	3.081%	3.318%	6.872%
10	0.079%	5.371%	5.450%	0.000%	3.949%	3.949%	9.400%
11	0.158%	4.660%	4.818%	0.079%	3.870%	3.949%	8.768%
12	0.158%	5.687%	5.845%	0.553%	4.423%	4.976%	10.821%
13	0.553%	3.555%	4.107%	0.474%	4.265%	4.739%	8.847%
14	0.711%	2.923%	3.633%	1.501%	3.397%	4.897%	8.531%
15	0.948%	2.607%	3.555%	1.817%	2.528%	4.344%	7.899%
16	1.106%	2.923%	4.028%	1.975%	1.422%	3.397%	7.425%
17	2.607%	1.896%	4.502%	2.054%	1.738%	3.791%	8.294%
Grand Total	7.188%	#######	51.185%	######	#######	48.815%	100.000%

Task 2.2:

Visualisation: For easy understanding we visualise the distribution into bar charts where we insert bar chart for our table.



The bar chart showcase gender wise schooling of children between age of 6 to 17. Peak of Schooling age for both group is 12 suggesting the most attendance at certain age.

1. Gender Difference: The chart shows gender disparting for schooling status. Where the yellow bar suggesting male population in school is slightly more than the girls (shown in orange bar).
2. Dropout Trend: For age group of younger children (6-12) vast majority is currently schooling. However, after the age of 13 there's steady increase in children dropping out. It is clear for both genders.

So, constructing policies considering these two issues would be beneficial.

Task 2.3 Probability:

a. The probability that a randomly selected schooling-age boy is currently out of school is 19.09%.
b. The probability that a randomly selected schooling-age girl is currently out of school is 14.04%.
c. No, boys are more likely to be out of school than girls.
d. The probability that a randomly selected primary school-age child (age 6 to 11) is currently out of school is 2.054%.
e. The probability that a randomly selected lower secondary school-age child (age 12 to 14) is currently out of school is 3.95%.
f. The probability that a randomly selected upper secondary school-age child (age 15 to 17) is currently out of school is 10.51%.
g. Yes, older children more likely to be out of school than younger children.

Task 3: Distribution of education related expenditure

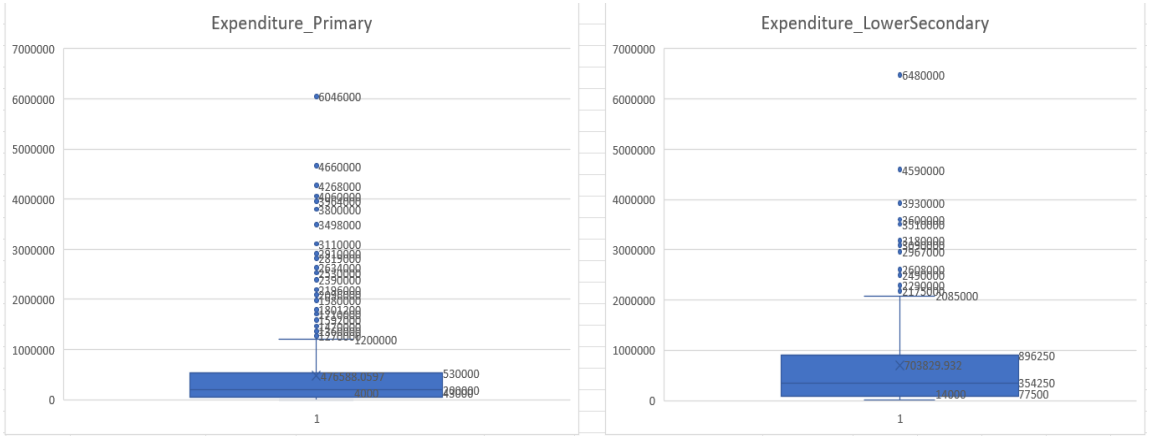
In earlier stage we saw the trend of children quitting schools after age 12. To solve this problem, we need to look further into the cause. One of the reasons might be expenditure on education. Let's do analysis on how it affects children's schooling rate.

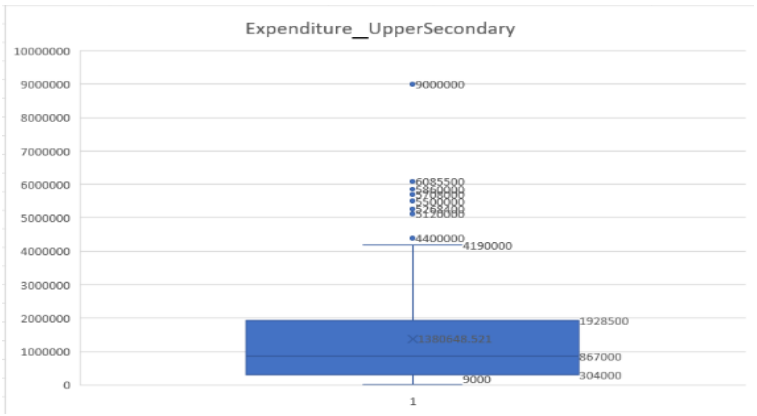
Task 3.2:

a.) Descriptive Statistics: To understand expenditure of each age group we performed descriptive analytics on those three-age variable. For that we made table n perform descriptive analytics from data analysis add-ins of excel. We got the below results.

Expenditure_Primary		*Expenditure_LowerSecondary*		*Expenditure_UpperSecondary*	
Mean	476588.0597	Mean	703829.932	Mean	1380648.521
Standard Error	33865.89221	Standard Error	55696.19707	Standard Error	115608.2777
Median	200000	Median	354250	Median	867000
Mode	25000	Mode	40000	Mode	680000
Standard Deviation	784052.0895	Standard Deviation	954990.844	Standard Deviation	1502907.61
Sample Variance	6.14738E+11	Sample Variance	9.12008E+11	Sample Variance	2.25873E+12
Kurtosis	15.04844381	Kurtosis	7.455624561	Kurtosis	4.551662999
Skewness	3.434942337	Skewness	2.474167149	Skewness	1.918958326
Range	6042000	Range	6466000	Range	8991000
Minimum	4000	Minimum	14000	Minimum	9000
Maximum	6046000	Maximum	6480000	Maximum	9000000
Sum	255451200	Sum	206926000	Sum	233329600
Count	536	Count	294	Count	169
IQR (Q3-Q1)	487000	IQR (Q3-Q1)	813000	IQR (Q3-Q1)	1605000
Q3	530000	Q3	891500	Q3	1913000
Q1	43000	Q1	78500	Q1	308000

b.) Visualisation: To better understand this statistical table, we created box-whisker plot for each variable.





C) Discussion:

1.) Central Location: Central location for each variable is measured by mean. Here Expenditure_UpperSecondary has the highest amount of mean (1380648.521) followed by Expenditure_LowerSecondary(703829.923) and Expenditure_Primary(476588.0597). even with median this trend is quietly prominent. This suggests that on average money spent on Upper secondary year is more than secondary and primary years.

2.) Symmetry: For all three variables mean is greater than median. Which shows the positive skewed nature. Primary has highest skewness (3.434) followed by secondary (2.747) and upper secondary (1.918). It means majorly primary expenditure is relatively lower than secondary and upper secondary's expenditure but from the whisker box plot you can see even in primary expenditure has some outlier.

3.) Spread: The spread data measured by IQR and standard deviation also differs across various age group. Expenditure_UpperSecondary has highest standard deviation (1502907.61) and IQR (1605000)

showing the most variation in expanses followed by LowerSecondary (SD=954990.844 & IQR=813000) and Primary (S.D.=784052.0895 & IQR=487000). Suggesting Primary has a smaller number of variations. Even in the whisker-box plot it is suggested that Expenditure_UpperSecondary covers broader range than other two.

Task 3.3:
A detailed analysis about whether the average educational expenditure per upper-secondary school child exceeds 1,380,000 riels per academic year in Cambodia, using a 5% level of significance.

1. Setting Up the Hypotheses:
Null Hypothesis (H_0): $\mu \leq 1{,}380{,}000$
Alternative Hypothesis (H_1): $\mu > 1{,}380{,}000$

2. Significance Level (α): = 0.05 (5%)

The appropriate test statistics is because we don't know the population standard deviation.

3. Test Statistic (t-value):
$t = (\bar{x} - \mu) / (SD / \sqrt{n})$
$t = (1{,}380{,}648.52 - 1{,}380{,}000) / (1{,}502{,}907.61 / \sqrt{169})$
Tobs = 0.005609639 and Tcritical = 1.653974208

4. P-value: 0.497765417

5. Confidence Interval
Lower Bound = 1189435.411

8. Decision Rule:
- P-value Approach:
  - If p-value $\leq \alpha$, reject the null hypothesis.
  - If p-value $> \alpha$, fail to reject the null hypothesis.
- Test Statistic Approach:
  - If the calculated t-statistic > critical t-value, reject the null hypothesis.
  - If the calculated t-statistic ≤ critical t-value, fail to reject the null hypothesis.
- Lower Bound Test
  - If the Lower bound for the one-sided test is > Target (1,380,000), reject the null hypothesis.
  - If the Lower bound for the one-sided test is < Target (1,380,000), fail to reject the null hypothesis.

9. Results:

We fail to reject the null hypothesis at 5% significance level based on *all three-methods- T* statistic Comparison, P Value, and Confidence Interval.

- T Test, Since Tobs < Tcritical, we fail to reject the null hypothesis.
- P Value, Since p-value is > alpha we fail to reject the null hypothesis. There is not enough evidence to conclude that the average expenditure is greater than 1,380,000 riels.
- Lower Bound: Because Lower Bound < Target, we fail to reject the null hypothesis.
- There is *insufficient evidence* to support the claim that the average educational expenditure per upper-secondary school child in Cambodia exceeds 1,380,000 riels per academic year.
- While the sample mean (1,380,648.52 riels) is slightly above the hypothesized value (1,380,000 riels), the relatively high variability in the data (indicated by the large standard deviation of 1,502,907.61 riels) does not let us conclude that this difference is statistically significant.

Task 4: Cause of Expenditure

In earlier task we performed statistical analysis on expenditure based on age group in this task we are going to focus on factors of those expenditure.

Task 4.3

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-567239.2544	173696.8321	-3.26568566	0.001129345
Age (Curated)	96957.17163	9624.007332	10.07451141	8.53246E-23
M (Curated)	-69574.3217	58864.90795	-1.18193206	0.237515618
K (Curated)	-55968.90872	106221.002	-0.52691001	0.598373769
Years (Curated)	87312.4497	7657.429402	11.4023186	2.17553E-28
Under 18 (Curated)	-83737.48635	23722.83008	-3.52982701	0.000434904

a.) Threshold for significant is given as 5%. According to that only **age, years and under 18** has lesser p-value than 0.05. So, only those three are statistically significant and not all. Household being Khmer and child being male factors are not statistically significant due to higher p value then 0.05.

b.) Age: The coefficient is 96597.17163. This means that in average for each one unit increase in age the dependent variable increased by 96597.17163(all other variables are constant). P-value suggests highly significant relationship. Years: The coefficient is 87312.4497. This means that on average for one additional year the dependant variable increased by 87737.45 units. P-value suggest highly significant relationship. Child being under 18: For each additional child there is under the age of 18, the educational expenditure decreases by 63127.405 Riels.

c.) Since the p-value is higher than the significance level of 0.05, there is no statistical significance and evidence that one gender is more biased than the other.

d.) Based on output there is no direct indication that it not ethnic biased.

Regression Statistics	
Multiple R	0.459587674
R Square	0.21122083
Adjusted R Square	0.207249133
Standard Error	926482.9065
Observations	999

e.) The difference between the value of R^2 and adjusted R^2 is very less and suggest that model is not overfitted.

f.) No. The model is not explaining variation of household expedition on children's education very well. The R^2 value is only 0.21122 which means that only 21.11% of variation in the model is explained by these variables remaining 79% expedition is still unexplained. Normally R^2 value near 1 tells model is better fitted and explained the variation of dependent variable. Here even adjusted R^2 value is 0.207249 more lesser than R^2 value suggest that model is not very much explanatory.

g.) Since p-value of the test statistic is less than 5% (significance level), therefore, there is insufficient statistical evidence to reject the null hypothesis of normal distribution of the variable. The Jarque-Bera test has determined that the distribution is not normally distributed. With such a large test statistic, it strongly suggests that the p-value would be very small (close to 0). As mentioned before, a p-value close to zero indicates very strong evidence against the null hypothesis.

h.) No, there isn't any other variable that can be considered affecting the child's education extensively. Some hypothetical variables (not given in the dataset) could have been head's income and household's debt. Head's income level would be directly proportional to the spending towards children's education and the the household's debt would be inversely proportional to the spending of the children's education.

TASK 5

Summary: Based on the preliminary analysis of the Cambodia Living Standards Measurement Survey (LSMS) 2019-2020 data, several key findings,

The data shows that diverse groups of people spend differently on their children's education. Some groups spend more than others, and this highlights the need to give extra financial help to families who have less money. There are also differences in spending based on gender and ethnicity, which could show that there are social or cultural factors making it harder for some groups to access education.

Policy Suggestions:

1. Officials should make policies that provide financial help like offering subsidies or some finance can help lower-income families afford education.
2. Government should place policy that promotes education equal for boys and girls which to make sure that both boys and girls have the same chance to access education.
3. To support minority communities like work on improving education access for ethnic groups that are underrepresented, or face barriers make policy or reservation in education sector.
4. Policymaker should explore innovative ways to help families to feel less burden to pay fees of children such as introducing loans policies which will be affected from certain age.

Insights and Recommendations:

After analysing the dataset, we found that head of household and related factors will impact the educational expenditure. However, this is not the only relationship that may be considered as the sample set suffers from specification errors. Therefore, to make an informed judgement further data and considerations need to be investigated, which this data set could only be used as a minor and negligible preliminary insight. The analysis shows us that the key factors that affect how much families spend on education. Overall, we have concluded that this preliminary data has limited insights as there are better ways to determine educational expenditures. To make predictions more accurate in the future, models should consider other factors like the parents' education levels, salaries and government support. Also consider COVID-19 effects in longer planning. Using data to guide policy decisions will help make education more accessible and fairer for everyone.