# Final report

**Group 12: Zohd Khan, Siddarth Vinnakota, Devansh Vora, Tracy Zhu**

## 1. Introduction:

Situated between its larger neighbors of India and China lies the small nation of Nepal. It is a beautiful landlocked country with mountains everywhere. It is within this location that roughly 7 out of the 9 tallest peaks on the planet find their place. Nepal, although being a small country, can be representative of living conditions in South East Asian countries as most of them have the same values. Nepal along with India and China is a developing country with a lot of socio-economic challenges.

*Questions of interest:*

- What are the relationships between the members per sleeping room proportion between the socioeconomic factors of food insecurity, education level, and ecological region.
- With these factors, can we create a model that identifies if a household is overcrowded or not? As a follow up, what factors affect this classification the most?

*Why Should We Care?*

As mentioned above overcrowding is an issue that is being faced by many countries situated in Southeast Asia. Nepal has almost 6 times more people living per mile squared compared to the United States of America. This is a significant number, and this could happen due to a variety of issues. One of them is the nature of lying with parents in Nepal, wherein people tend to live with their parents and grandparents, and Nepal households seem to be overcrowded to a person belonging to the West, but to a citizen of India, it might seem normal.

Provided this condition it becomes vital for us to do an analysis where we consider some of the major factors that might affect the well-being of the people of Nepal.

*Primary Variable of Interest:*

Our primary variable of interest is the “Number of People Sleeping per Bedroom”. This variable serves an important role in determining the household condition for Nepal. This variable was not directly available to us in the dataset thus we had to divide the total number of rooms used for sleeping in a household by the number of people living in the house.

*Theory and Hypothesis:*

The framework of this project is built on the assumption that overcrowded living conditions in Nepal can have substantial effects on variables like food insecurity and wealth management within households. We also wanted to determine if there are crowded households in the hilly mountain region where the natives are expected to live. The hypothesis assumes that the increase in the Number of People Sleeping per Bedroom is likely to correlate with an increased number of people with food insecurity, given the potential strain on resources and limited space for food storage.

Another important variable we wanted to focus on was the effect of our primary variable of interest on Education. Many factors limit Education for children of certain households. Limited number of resources being the primary one, we wanted to examine if the number of people

sleeping in one room correlated with the fact that a household has a limited number of resources to support education costs for children.

*Existing Knowledge:*
The current existing data on Southeast Asian countries examines how overcrowded living conditions of a country affect well-being. The goal of this project is to add a new dimension to the broader knowledge base by focusing on the unique variable, the Number of People Sleeping in Each Bedroom which focuses on the overcrowded living conditions of a household. The goal is to gain a deeper understanding of how living conditions and key well-being indicators interact.

## 2. Literature Review:

As mentioned above, most of the previous projects have been based upon the analysis of crowded conditions of a country and how it is performing. There is little to no research done for a very narrowed-down variable which is a key factor in calculating the overcrowding of a household. It provides a deep understanding of how an overcrowded house can differ from a house that does not have overcrowding. The gap we want to fill by doing this project is to try and convey that internally in a country many things decide the fate of people, but some of them are neglected easily like the overcrowding of houses.

## 3. Materials and methods

*Data:*
We acquired the data from the DHS database available at https://dhsprogram.com/data/. We had to request access to the data we wanted to work with and state the reason for acquiring the data. After getting the data we did a preliminary Exploration of the dataset. The dataset had 13786 rows and 3027 columns. The columns consisted of the number of rooms used for sleeping and the total number of people in the household were named as hv009 and hv216. The dataset also contains a lot of information about the household including health and wealth.

Throughout this report, we will sometimes refer to different factors by their id names. What each id represents can be found below:

- Hv009 - the total number of members in a household
- Hv216 - the total number of rooms primarily used for sleeping
- RoomProp - the proportion of the number of household members and the total number of sleeping rooms. This value is obtained by dividing hv009 and hv216
- Shecoreg - the ecological region the household is located in. There are three regions: mountain, hill, and terai
- Hfs_mod - the probability of moderate food security, measured as a numeric value
- Hfs1 - a categorical variable that asks if households worry about the lack of food. This is used to calculate hfs_mod
- Hv106_01 - a categorical variable representing the highest education level attained per household

- Overcrowded - a categorical variable that determines if a household is considered overcrowded or not. This is based on our RoomProp variable. For the purpose of this report, our threshold will be 2.

## 4. Methods and Results

***Correlation between Room Proportion and Food security:***
In our investigation of the socio-economic factors affecting the well-being of Nepalese households, we focused on the relationship between household overcrowding and food insecurity. We utilized the comprehensive Demographic and Health Surveys (DHS) Program database, which provided us with a dataset including variables such as 'hv009' (number of people per household), 'hv216' (number of rooms used for sleeping), 'hfs_mod' (probability of moderate or severe food insecurity), and 'hfs1' (household worry about food due to lack of money).

To calculate our primary variable of interest, 'RoomProp, ' we calculated a ratio of 'hv009' to 'hv216'. Our analysis began with assessing the link between 'RoomProp' and 'hfs_mod', a numerical proxy for the likelihood of food insecurity. However, recognizing the importance of subjective experiences in the understanding of food security, we also closely examined 'hfs1', a categorical variable reflecting households' concerns about food access due to financial constraints, with 'no', 'yes', 'refused to answer', and 'don't know' as possible responses.

Our analytical methods included Pearson and Spearman correlation coefficients to identify linear and monotonic relationships, respectively, and a t-test to compare the means of 'RoomProp' between different groups within 'hfs1'. We cleaned the dataset by removing missing, infinite, and outlier values, ensuring a robust and reliable analysis.

The correlation analysis using 'hfs_mod' initially revealed a weak but statistically significant relationship with 'RoomProp', with a Pearson correlation coefficient of 0.174 and a Spearman rank correlation of 0.183. These figures indicated a marginal tendency for households with more individuals per bedroom to experience higher probabilities of food insecurity. However, the weak strength of these correlations suggested that overcrowding might not be the most significant factor influencing food security.

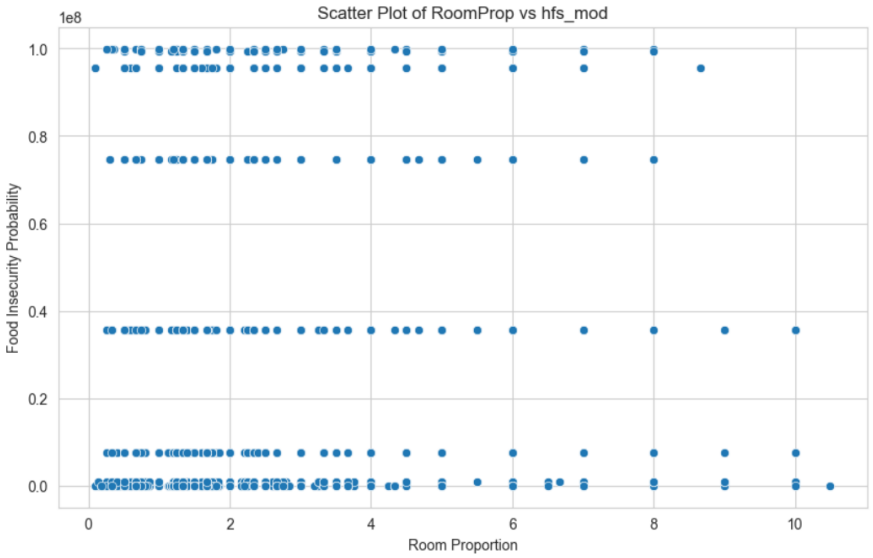
Upon removing outliers, the Pearson and Spearman correlation coefficients slightly decreased to 0.137 and 0.153, respectively. This further supports the hypothesis that while a relationship exists, it is not strong. The presence of outliers could have exaggerated the initial relationship.

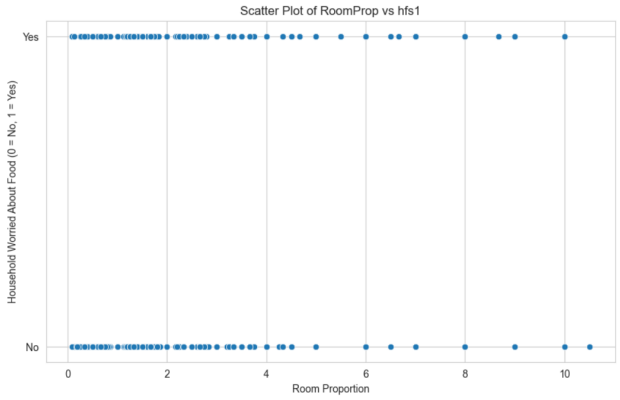
Considering the modest correlation with 'hfs_mod', we examined 'hfs1' to ensure we captured the multifaceted nature of food insecurity, including subjective experiences. The Spearman rank correlation between 'hfs1' and 'RoomProp' was 0.159, supporting a slight association. Importantly, the t-test conducted on 'hfs1' provided a more pronounced result, yielding a t-statistic of -20.166 and a p-value effectively at zero, which indicated a statistically significant difference in room proportions between households with and without food insecurity concerns. This discrepancy in findings between 'hfs_mod' and 'hfs1' underscored the necessity of including

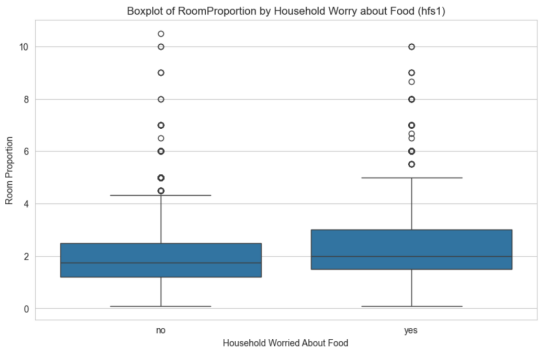
both objective and subjective measures in our analysis to obtain a comprehensive understanding of the impact of overcrowding on food insecurity.

Visual analysis through scatter plots and boxplots corroborates these findings. The scatter plots show a dispersion of data points that do not cluster tightly around any particular trend line, indicating a weak correlation. The boxplot comparing 'RoomProp' by household worry about food shows overlapping interquartile ranges, reinforcing the statistical analysis that the relationship between these variables is not strong.

These findings, taken together with the correlation results, paint a nuanced picture of the relationship between overcrowding and food insecurity in Nepal. While there is some association between more people sleeping per bedroom and increased food insecurity, the relationship is not strong, and the significant difference in means as indicated by the t-test complicates the narrative. This complexity underscores the need for multifaceted approaches to address food insecurity, which should take into account not only housing conditions but also a variety of other socio-economic and cultural factors.



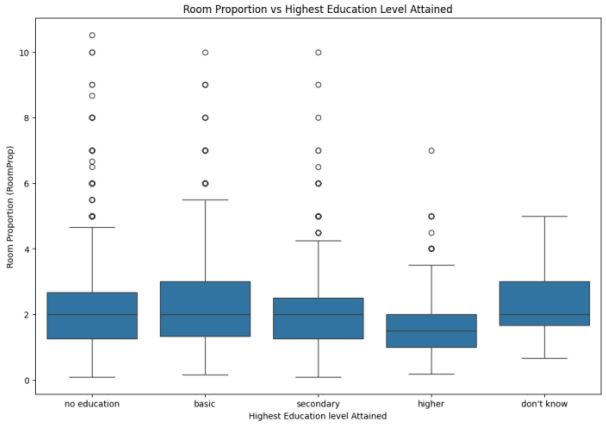



***Correlation between Room Proportion and Education:***

To better understand how significant our primary variable of interest is, we also examine its relationship with education levels of those in Nepal, specifically by comparing it with the "Highest Education Level Attained" variable (hv. This variable, which contains the categories "No education/Preschool", "Basic", "Secondary", "Higher", and "Don't know", serves as a general representation of how educated the respondents in the survey are.

We first create a boxplot through Python to visualize the relationship between our variable of interest and the education variable. Based on the plot (depicted below), it is clear that the lines in each of the respective boxes are slightly different between categories. This serves as one indication that if the variables do have an association, it is fairly weak, because the fairly similar medians between groups suggests there is not much difference. This idea is further supported by the fact that the actual boxes, which represent the respective Interquartile Ranges (IQRs) for each category, are also fairly similar to one another. The visualization approach alone does not offer us the basis to form a conclusion as we must determine where the most significant differences are between groups, and how significant those differences truly are.




```
ANOVA test between RoomProp and hv106_01: F-statistic = 31.836450844648734, p-value = 1.9097802794526754e-26
```

We now conduct an ANOVA test in Python to further analyze the relationship between these two variables. The F-statistic of approximately 31.8364 (depicted above) tells us the ratio of the variance between education level groups to the variance within each group. The relatively high value indicates there is a greater degree of separation between the group means. This, along with the very small p-value (depicted above), suggests a statistically significant association between the education level and room proportion. This is because the ANOVA points to real differences in the means for the different groups. Although the ANOVA does suggest a significant association between the variables, we must perform further analysis to determine in which specific groups the differences lie so that we can determine why exactly there is an association between variables.

| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|---|---|---|---|---|---|---|
| basic | don't know | 0.1637 | 0.9599 | -0.4886 | 0.8159 | False |
| basic | higher | -0.4729 | 0.0 | -0.6195 | -0.3263 | True |
| basic | no education/preschool | -0.0457 | 0.2999 | -0.1102 | 0.0188 | False |
| basic | secondary | -0.2164 | 0.0 | -0.2907 | -0.1422 | True |
| don't know | higher | -0.6365 | 0.0687 | -1.302 | 0.0289 | False |
| don't know | no education/preschool | -0.2094 | 0.9059 | -0.8617 | 0.4429 | False |
| don't know | secondary | -0.3801 | 0.5057 | -1.0334 | 0.2733 | False |
| higher | no education/preschool | 0.4272 | 0.0 | 0.2802 | 0.5741 | True |
| higher | secondary | 0.2564 | 0.0 | 0.105 | 0.4079 | True |
| no education/preschool | secondary | -0.1707 | 0.0 | -0.2457 | -0.0958 | True |

We proceed by performing a Tukey's HSD (Honestly Significant Difference) test to determine where in which groups the mean differences lie and extreme those differences are. Based on the results (depicted above), we can conclude differences between means in certain groups are of substantial size as well as statistically significant. For example, the comparison between "basic" and "higher" education categories points to substantial differences given the "meandiff" value of -0.4729, which suggests that the mean room proportion is greater for the "higher" education group than it is for the "basic" education group. We can also verify that this difference is statistically different given the p-adj value of 0 and value of "True" in the reject value, which means we can reject the null hypothesis that there are no no differences between the means of the two groups.

We can also conclude that the differences between basic and secondary education categories are of substantial size and statistically significant based on the adjusted p-value and rejection decision for that comparison.

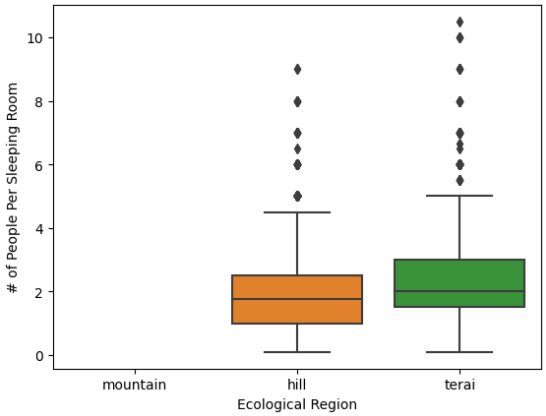
The differences between groups for basic - secondary, higher-no education, higher-secondary, and no education - secondary comparisons are also of substantial size and statistically significant based on their respective results in the Tukey's HSD test. The most notable conclusion from these comparisons is that given mean values of .4272 and .2564, the respondents in the "higher" education category have fairly large mean room proportion values compared to the "no education/preschool" and "secondary" education categories, respectively. We can also confidently say that those in the "secondary" category have higher mean room proportion values compared to those in the "no education/preschool" category.

Although there are some comparisons where the differences are not statistically significant, such as the comparison between "basic" and "no education/preschool" education categories, the several comparisons leading to the opposite conclusion point to the overarching idea that there is an association between our Room Proportion variable and the chosen education variable representing education. Specifically, the association is that respondents with higher education levels generally had higher room proportion values.

***Correlation between RoomProp and Ecological Region***

A major change with the Ecological Region was that out of the 3 classes, we removed the mountain class for this experimentation. This is because, compared to the other two classes, there were significantly less number of households located in the mountainous ecological region. So, we instead take a look at the remaining two classes.

After removing the outliers, below are the associated boxplots of both of the classes:




Based on this graph, we can see a noticeable trend: households in the terai region on average have a higher room proportion than households in the hill region. However, the next step is to identify if this change is significant enough to call it a trend. We will utilize a statistical test to determine this. In this scenario, two tests are used: the T-Test and the Mann-Whitney Test. Both of these tests are similar in that the main goal is to take a null hypothesis that the means are the same, and either reject or fail to reject it. The main difference between the two tests is that the T-Test is parametric and requires the assumptions of a normal distribution and equal variances, while the other test is nonparametric and does not require those assumptions. To be safe, we employed both of these methods, and below are our associated results.

| | T-Test | Mann-Whatney |
|---|---|---|
| Test Statistic | -12.08 | 16342883.5 |
| p-value | 2.14E-33 | 5.04E-37 |

The most noteworthy value we will be looking at is their respective p-values. Because both the p-values are extremely small (smaller than any regular significance level), we can safely reject the null hypothesis, and conclude that this difference in the average room proportion for both of these classes is statistically significant, implying that this feature is an important factor to consider.

***Creating a Combined Model***

With these factors, we next want to see if these factors can be used to create a classification model that can determine if a household is considered “overcrowded” or not.

Firstly, we have to define what is the threshold for “overcrowded”, as there is not much of a definitive number when it comes to Nepal. In the US, most consensus suggests that the threshold of when a household is considered overcrowded is when the number is between 1-1.5 people per room. For the sake of this model, we will be using a threshold of 2, as that provides a similar quantity of those below and above the threshold. To clarify, any households with a RoomProp less than 2 is not considered overcrowded, while any households greater than or equal to 2 will be considered overcrowded.

The first model constructed was a Categorical Naive Bayes Model, which takes into account the factors of ecological region, education, and food insecurity. However, one of the main assumptions of this process is that the features are independent, so in order to account for that assumption, we also do the same classification process with similar accuracies with both the Support Vector Classifier (both linear and rbf variant) as well as the Random Forest Classifier. A support vector classifier works by finding the hyperplane that maximally separates different classes by identifying support vectors, which are data points crucial for determining the decision boundary, and a random forest classifier is an ensemble model that uses a collection of random decision trees that provides a classification by traversing through the features. Below are the accuracy reports of each model. Additionally, we ran a cross validation grid search in order to fine tune the model to find the best version of the model. This meant creating a new model with slight variances and finding the variation with the best accuracy.

| | Naive Bayes | SVC (linear) | SVC (rbf) | Random Forest |
|---|---|---|---|---|
| Accuracy | 0.58 | 0.57 | 0.56 | 0.56 |

As shown in the table above, each model is able to correctly classify around 60% of the data as overcrowded or not. While it is decent, it is not ideal, which we would classify as having a 70% accuracy or above. Unless the model can be improved, these 3 factors in food insecurity, education level, and ecological region, cannot be primarily used to classify the overcrowdedness.

The next step was to look for the most important features of the model, and below is a table of the features, as well as their level of importance. This is extracted from the Random Forest Classifier

| Feature | Importance |
|---|---|
| Hfs1 (food insecurity) | 0.467413 |
| Shecoreg (ecological region) | 0.360393 |
| Hv106_01 (highest level of education attained) | 0.172194 |

As seen from this table, food insecurity and ecological regions have the highest factor in classifying the model, while the education level is very ineffective. This shows that at least in this model, the highest level of education has nothing to do with the room proportion of Nepali Households.

## 5. Discussion of Findings:

Using the number of people sleeping in each bedroom in Nepal, we discovered some surprising results when analyzing this data alongside the food insecurity variable. As we anticipated, there should have been a strong correlation between the two variables; however, it appears that the relationship between overcrowded living conditions and food insecurity is weaker than we expected. Moreover, this fits well with the existing literature on family values that exists in Southeast Asian countries, where most people live in joint families for generations and it is considered rude to separate oneself from a family.

These results are consistent with the theories and data that currently exist, which show how strongly Nepali people value their families. Living in a shared space is not associated with food insecurity, which is consistent with new research highlighting the relationship between housing, financial standing, and the ability to get necessities. These findings not only add to the gap in literature but also offer specific insights into the Nepalese reality.

In regards to the correlation between our variable of interest and education, although we found that there was a decent correlation between the education of a household and how crowded that household is, the direction in which this correlation manifests is opposite to what we expected. Contrary to typical socioeconomic patterns, our analysis suggests that households with higher education levels are associated with a greater number of individuals per room. This is an unexpected conclusion given that the societal expectation would be to have less people per room as a household has a higher education level, and that those with higher education would be expected to have higher income, and therefore be able to afford more space for each individual person. It is important to understand that while resources for education might be limited to a household due to a variety of factors, our results suggest that overcrowding is not one of them. This could be a breakthrough as most people have a misconception that if a household is overcrowded, they would limit their children going to school, but contradictorily they don't. This could also be due to the ever-increasing literacy rate in Nepal. Literacy in Nepal grew by about 11.52% from 2011 to 2022. It's conceivable that those with higher education in Nepal may choose to live with a greater number of people in their household, potentially overestimating the number of individuals their living space can comfortably accommodate. Alternatively, it may be that individuals with higher education levels opt to live in larger family units if they can afford it, despite the potential for overcrowding per room. This tendency could be illustrative of cultural norms in Nepal, where communal living and extended family households are valued and prevalent, suggesting that the measure of room proportion does not solely reflect economic capability but also cultural preferences and social structures.

When it comes to ecological regions, there seems to be a difference in average proportion between the hill and terai regions. More specifically, the terai region seems to have a higher proportion. The reason why this statement is significant is because it provides an overview of the internal factors that could be attributed to RoomProp. For example, the Hill terrain is rugged, while the Terai region is flatter. Additionally, the Terai region is considered more developed, and the land is more fertile. Additionally, the population percentages remain relatively the same. By identifying there is a proportion difference between the two regions, we can further develop this finding by looking at these more internal differences to find the clearest attribute to this average difference. Future work would delve into this, by looking at the more individual factors and attributing them to the overall model. For example, we may want to look at the individual materials that make up a house, because there might be some material that we can use to determine the

The classifier with the 3 factors proved as an inconclusive classifier on whether or not a household can be considered overcrowded or not. However, based on the individual correlation being present, albeit weak, there is a possibility of further improving a possible model that can help classify that. For example, we could increase the number of features, to provide more possibilities and directions that the model could use to classify the values. Some features we can look at are the material makeup of the homes in each house, as not only they can represent

wealth, but there may be heat-related properties based on the material used. However, there are a wide variety of classes, each with a varying amount, from 6000 of one class to under 500, so either more data would need to be collected, or we would have to drop some classes. Additionally, there could be more complex models such as neural networks and decision trees that can yield us a better conclusion. Finally, while not currently possible, we could increase the number of households taken, or even look into households in the previous years. Speaking of previous years, a potential outlook we could look at is analyzing how the average room proportion changes over time, to identify any major trends or events in Nepali history that could be seen as a certain factor (for example, if we notice a change in the room proportion during the COVID years, does that indicate that the presence of COVID is a factor or not). Ultimately there are many directions that we could go with our experimentation, but establishing a classifier would be beneficial for governments and other groups to identify what factors affect this possible overcrowding and address them accordingly.

***Real-World Implications and Limitations:***

Our work has implications beyond just the theoretical sphere of policy analysis; the implications of it extend to the practical realm as well. Using these findings, policymakers in Nepal could formulate interventions to address the challenges associated with food security issues, and educational opportunities that arise on an ongoing basis. Providing targeted assistance to economically disadvantaged households as well as initiatives to improve housing conditions will probably be instrumental in breaking the cycle of poverty. The limitations of this study must be acknowledged, however, as we cannot establish causal relationships based on the cross-sectional nature of the data and we do believe these conclusions can be generalized beyond the specific Nepali context in which they were conducted. Furthermore, it should be noted that while the "Number of People Sleeping per Bedroom" provides a unique perspective on living conditions, it does not contain all the facts about economic factors that affect living conditions.

Link to the github code: https://github.com/sidvin101/Sta160FinalProject