Statistical Analysis of Wealth Data for Households

Sanap Madhav

Department of Statistics

Savitribai Phule Pune University

April 1, 2019

Contents

1 Data Cleaning and Transformation

1.1 Data Cleaning

Observing data it can be seen that variable 'toilet share' has more than 2 lakhs missing values. Hence, this variable is not considered for further analysis.

There are 24,631 observations(i.e., households) whose response is 'other' for some variables, which doesn't give information about what type of facility is available at that household for that particular variable. Hence, these observations are discarded from further analysis. Further analysis is done on the basis of remaining 5,77,492 observations and 36 variables.

1.2 Data Processing

As response of variables like water, toilet, floor, wall, fuel and roof have more than two levels and that are nominal in nature (i.e., level 22 is not necessarily greater than level 11), they are converted to binary variables by using following logic. Amenities or facilities which are more expensive/represent high wealth of household labelled as one and zero otherwise.

- **For variable water:** Piped into dwelling, Piped to yard, Tube well or borehole, Protected well, Bottled water, Community RO Plant are more expensive drinking water sources and hence labelled as one. Further remaining responses/levels of this variable are labelled as zero. (Assuming that source of water is owned by household and it reflects high wealth of household up to some extent.)

- **For variable toilet:** Toilet facilities which require more money to build are labelled as one. That involves All FLUSH TOILET, Ventilated Improved Pit latrine (VIP), Composting toilet and Dry toilet. NO FACILITY and open pit are labelled as zero.

- **For variable floor:** Natural floors labelled as value zero. Further RUDIMENTARY and FINISHED floor are labelled as one.

- **For variable wall:** FINISHED walls are labelled one as response where Natural and RUDIMENTARY wall are labelled as response zero.
- **For variable fuel:** Even though natural gas, electricity and LPG are not much expensive these resources are mostly used by wealthy households. Whereas oil, coal, wood, animal dung and grass are used by poor households (assuming households have a free source of wood). Hence natural gas, electricity are labelled as one and oil, wood and coal are labelled as zero. No food cooked in house is labelled as one assuming person with this response takes outside food by paying money.
- **For variable roof:** All responses related to FINISHED roof are labelled as one and all other responses are labelled as zero.

2 Objectives

- To create a proxy variable to represent the wealth of households
- To report percentage of households having facility/amenity for each variable
- To visualize the data for better understanding
- To identify wealthiest households

3 Analysis

Figure 1 represent % of households with or without particular amenity. Here, we can observe that almost 90% of households have mobile, bed and bank account. Whereas proportion of households with servant, tractor or telephone (i.e., land line) is very less.

The actual percentage of having or not having certain amenity/service is reported in figure 2 and figure 3.

Here we can observe that mobile, bed, electricity and having bank account are common in most of the household whereas having car, computer, tractor, thresh and servant is rare for household in given data.

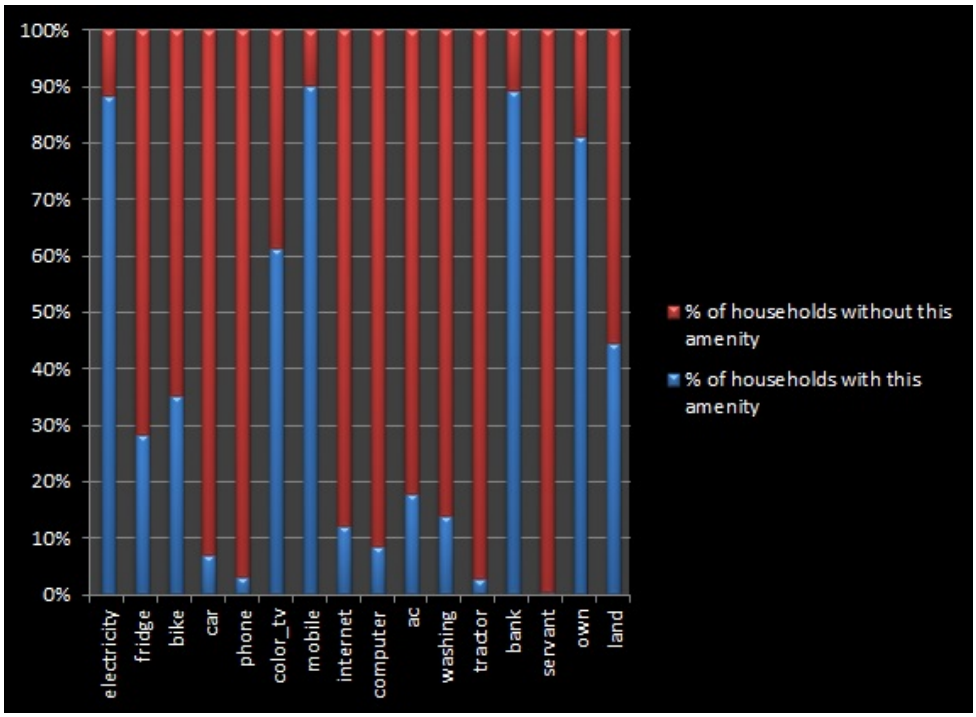


Figure 1: Percentage of households with or without certain amenities

Variable	% of households with this amenity	% of households without this amenity
mobile	89.99	10.01
bed	89.58	10.42
bank	89.31	10.69
electricity	88.24	11.76
roof	87.39	12.61
own	81	19
watch	77.23	22.77
chair	75.28	24.72
water	74.32	25.68
electric_fan	70.61	29.39
wall	69.28	30.72
mattress	68.67	31.33
floor	62.09	37.91
color_tv	61.16	38.84
pressure_cook	58.35	41.65
table	57.7	42.3
toilet	53.53	46.47
bicycle	48.53	51.47

Figure 2: Percentage of household with or without certain amenities for all variables

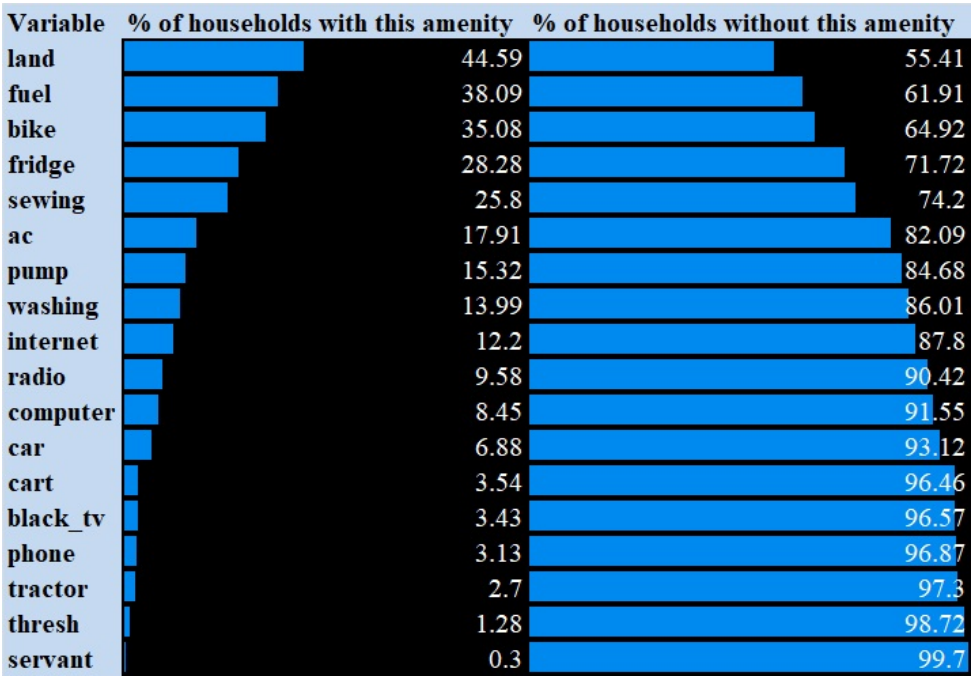


Figure 3: Percentage of household with or without certain amenities for all variables

3.1 Computation of score to represent the wealth of households

As having or not having certain amenity/service indicates wealth of household, adding variables can gives us an idea about wealth of that household. But, type of facility and price of that facility/amenity also matter. For example, having car and having mobile shows two different level of wealthiness. Hence, instead of simple addition we will add variables with specific weights.

These weights are decided on the basis of relative current price of amenities. Also using human mentality/logic of buying amenity for pride (means having some amenity treated as pride and consider that person as wealthy).

Weight of amenities in color pink are considered as 1/415, orange color represent weight 5/415, light green color represent 10/415, and dark green color represent amenities with weight 50/415.

For example: As fridge is approximately 5 times more expensive than electric fan (so we considered weight of fridge five times more than electric fan). Similarly car is approximately 10 times more expensive than fridge.

Using these weights, data is transformed by multiplying each variable with its corresponding weight (say weighted variables). Now, addition of these weighted variables gives us a required score. We treat these score as proxy variable to represent the wealth

Variable	Weight		Variable	Weight
car	0.1205		cart	0.012
ac	0.1205		pump	0.012
thresh	0.1205		servant	0.012
tractor	0.1205		electricity	0.0024
own	0.1205		radio	0.0024
land	0.1205		bicycle	0.0024
water	0.0241		phone	0.0024
toilet	0.0241		mattress	0.0024
floor	0.0241		pressure_cook	0.0024
wall	0.0241		chair	0.0024
roof	0.0241		bed	0.0024
fridge	0.012		table	0.0024
bike	0.012		electric_fan	0.0024
color_tv	0.012		black_tv	0.0024
mobile	0.012		sewing	0.0024
internet	0.012		watch	0.0024
computer	0.012		bank	0.0024
washing	0.012		fuel	0.0024

Table 1: Weights for variables

of households. This is logical because original variable takes value either zero or one for each household. Hence, original variable contribute only if that particular amenity is present (means only if original variable takes value as 1). Further, it's contribution is equal to it's weight.

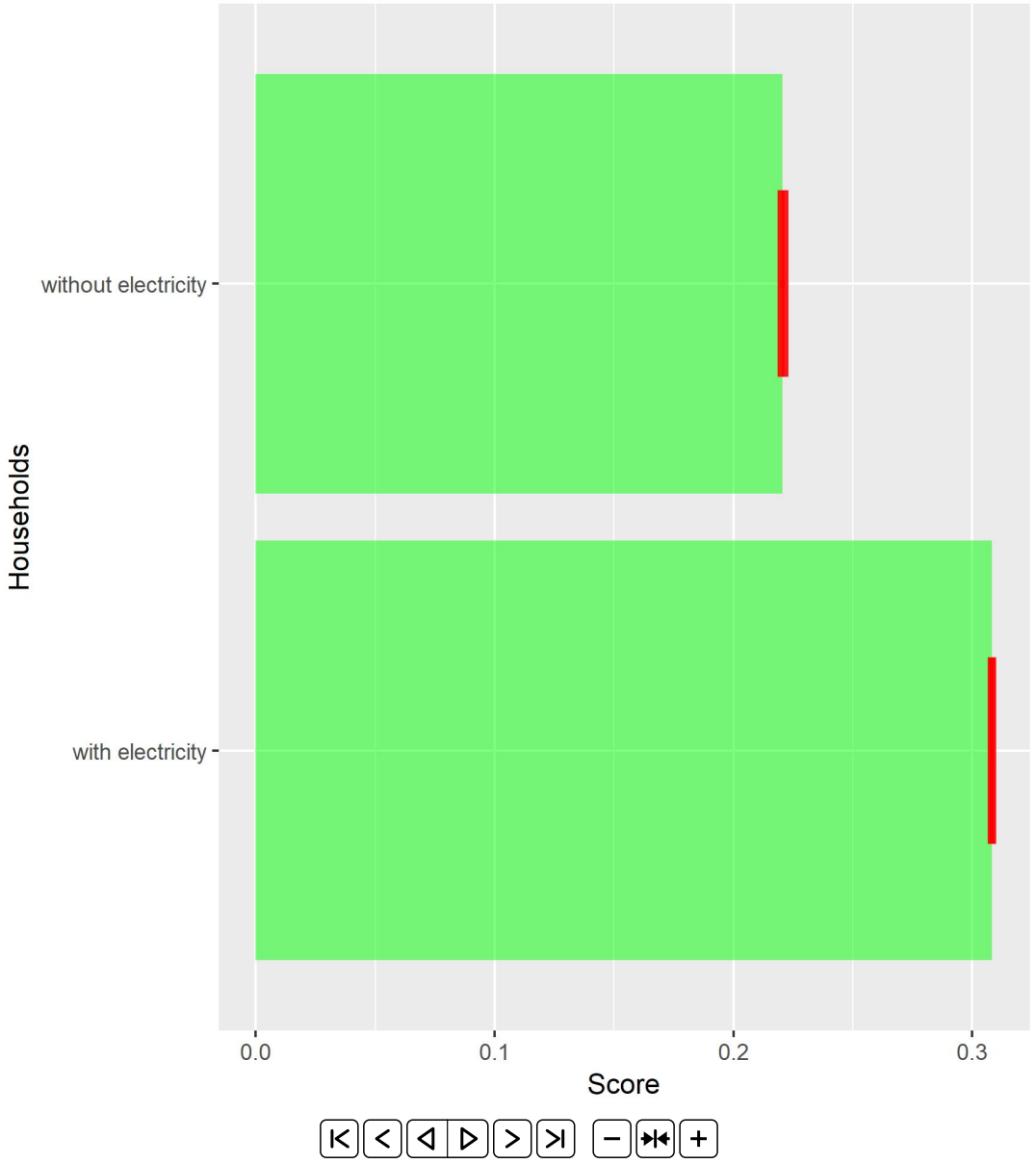
Even if all variable contribute, score (sum of weights) will take value one. Hence, the proxy variable (Wealth Index) will lie between zero and one. As having more amenities will increase the contribution of variables, high score indicates wealthy household whereas low score indicates poor household.

3.2 Comparison of household on the basis of presence/absence of amenities

Please wait and observe following animation

The animation below shows bar plot with error bar of new proxy variable (i.e., Wealth Index). Red line indicates variability (error) in average wealth score of households where green bar represent average score for household with or without given amenity

In almost all the cases error bar are not overlapping and hence average wealth score may differ significantly.



To confirm significant difference between score of variable 'wealth of a household' with and without amenity, z-test is used.

H_0: Average wealth index of household with and without certain amenity are equal.

H_1: Average wealth index of household with and without certain amenity are not equal.

Testing above hypothesis for each variable we get following p-values:

Variable	p-value	Bonferroni adjusted p-value	Variable	p-value	Bonferroni adjusted p-value
electricity	0	0	ac	0	0
radio	0	0	washing	0	0
fridge	0	0	watch	0	0
bicycle	0	0	cart	0	0
bike	0	0	pump	0	0
car	0	0	thresh	0	0
phone	0	0	tractor	0	0
mattress	0	0	bank	0	0
pressure_cook	0	0	own	0	0
chair	0	0	land	0	0
bed	0	0	toilet	0	0
table	0	0	floor	0	0
electric_fan	0	0	wall	0	0
color_tv	0	0	roof	0	0
sewing	0	0	fuel	0	0
mobile	0	0	servant	1.09E-297	3.93E-296
internet	0	0	water	2.02E-135	7.27E-134
computer	0	0	black_tv	1.47E-134	5.29E-133

Table 2: p-values

As all p-values are near to zero, we can conclude that there is significant difference between score (Wealth Index) of households with or without certain amenity

3.3 Identification of wealthiest households

By identifying outliers in 'Wealth Index', we get wealthiest households. Red color indicate wealthy household (whose score lies beyond cutoff Q_3+ 1.5*(Q_3-Q_1)) where Q_3 is third quartile and Q_1 is first quartile.

Here, wealth index is plotted for each household. We can observe layers in following figure which reflect different level of wealthiness in society.

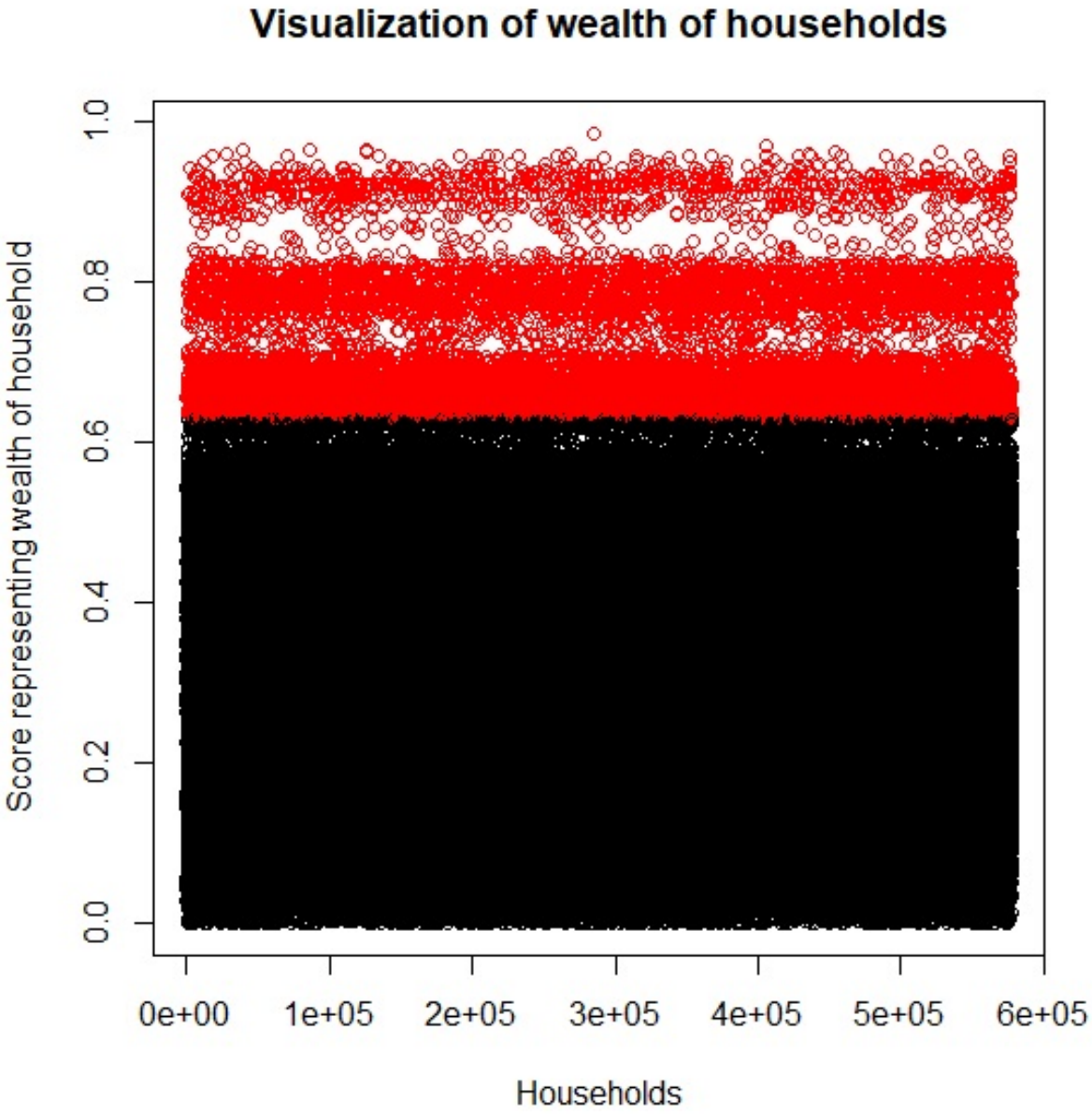


Figure 4: Score of households representing wealth

3.4 Graphical representation of variables representing basic facilities

Following are the pie chart indicating % of response in variable water sources, fuel use for cooking and toilet facility. This charts are drawn using original data of above variables where multiple levels are present.

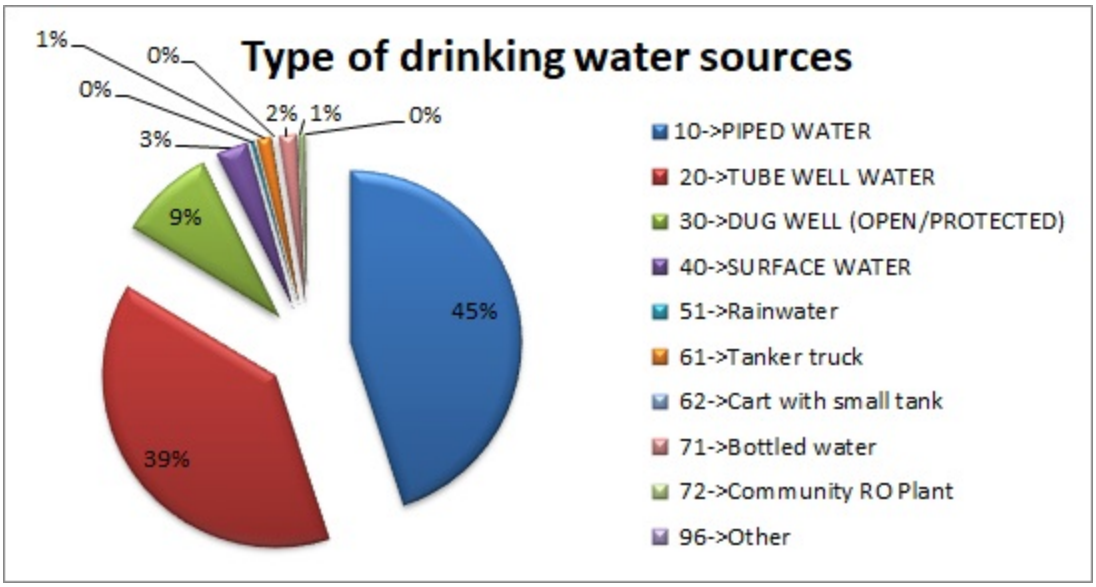


Figure 5: Sources of drinking water

Figure 5 indicates Piped water and Tube well water are most common resources for drinking water.

Figure 6 indicates LPG and Wood are most common fuel used for cooking.

Figure 7 indicates No facility and flush to septic tank are most common type of toilet facility.

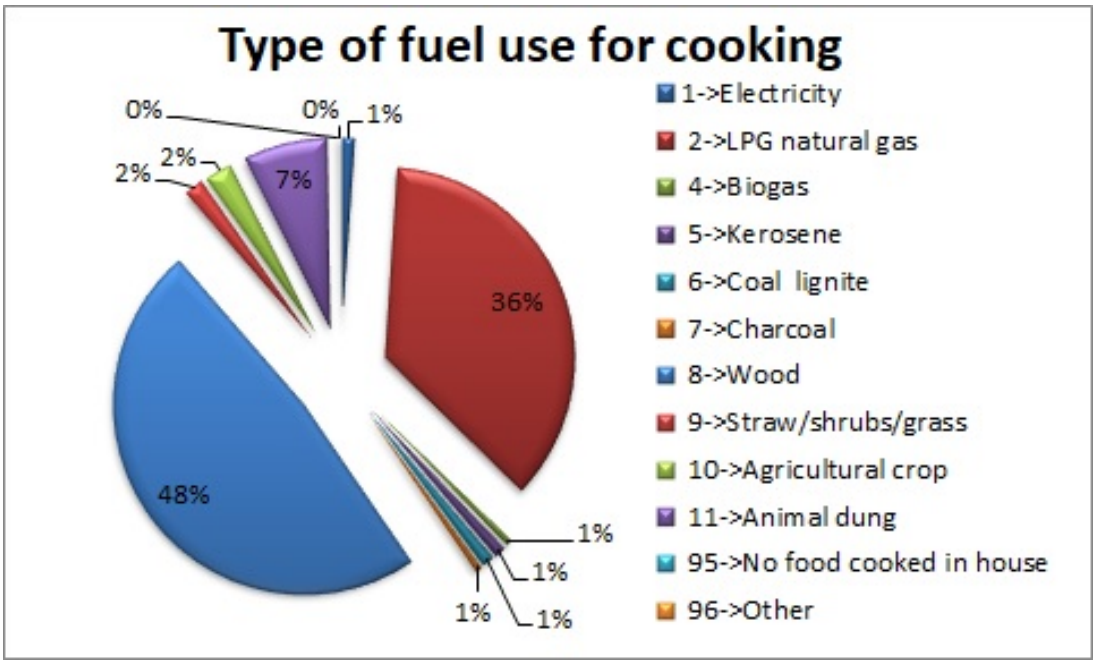


Figure 6: Fuel use for cooking

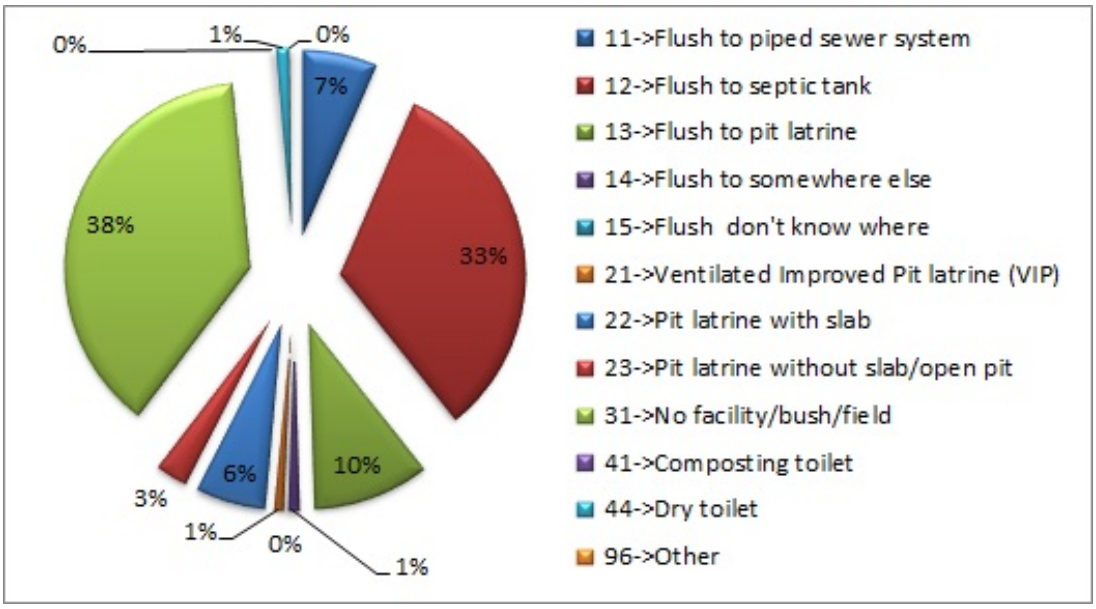


Figure 7: Type of toilet facility

Due to time constraint and computational limits, variables are transformed into two levels only which can easily extended to three or four levels, which may give a better approximation to proxy variable representing 'wealth of households'.