Story of A Slum: Pamplona Alta

Every year, in partnership with the Solidaridad en Marcha, Catholic churches throughout the city of Lima, Peru deliver thousands of Christmas gift boxes to the poorest residents in the city. The campaign, called Caja Del Amor, has been in operation for years and has subsequently built long-term and well-sustained relationships with many community leaders in these areas.

The list of gift recipients is built in coordination with the local community leaders, most of whom oversee around 150 families. These leaders choose the 5 to 10 families in most need of assistance in their respective community to receive the gift boxes.

La caja del amor
Un regalo de familia a familia

These networks served as a point of strength when we began to explore the creation of an in-depth survey to better understand the poorest urban populations in Peru. However, our focus needed to also be on action-ability of insight and the concrete opportunity of enacting positive change in the lives those we surveyed.

Therefore, we decided to focus on a region of high need, and an area where Solidaridad en Marcha had a significant footprint. This process of elimination led us to the region of Pamplona Alta near the San Juan de Miraflores municipality of Lima.

Pamplona Alta, A Brief History

Pamplona Alta is a shanty town riddled with extreme poverty, it has an absence of infrastructure, and a large portion of its community is without access to many basic human needs. The region has no public works, no paved roads, no public electricity, nor public access to water and sewage. If these services exist at all, they are provided by private companies at a premium price.

Water trucks provide the only (non-drinkable) water and they often cannot access many of the homes, especially in the upper portions of the region. Shallow latrines serve as a poor substitute for sewage and along with the pig farms, contribute to a high rate of parasitic infections, particularly among children.

The area was first populated in the 1990's as a result of a decade of terrorism that affected the entire country of Peru. During that period, populations from the surrounding regions began a mass migration to the outskirts of Lima, leaving the lives they knew to find safety in the proximity of the city. Many only spoke Quechua, few had employment for years after arrival, and none of these new residents owned the land on the edge of Lima where they would now call home.

In the opinion of many close to the matter, it has predominantly been this issue of land ownership that has ensured the continued impoverishment of the people of Pamplona Alta. Why land ownership? Well, owning the land is necessary, according to the government of Peru, for those in Pamplona Alta to receive basic municipal services. In order to attain roads, plumbing, water, schools and many other basic human needs, they must first own the land on which their homes sit.

The land is rocky, mountainous and un-arable, yet to the people of Pamplona Alta, it is home. Inside of this difficult landscape, there are two prevailing realities: those who squatted on public land and those who squatted on private land.

The valleys of this mountainous region were owned predominantly by pig farmers upon arrival, and even today many families live side by side to pigs in the lower part of Pamplona Alta. The owners of the land have lacked the resources to forcibly evict their unwanted tenants since their arrival, yet these tenants have now called Pamplona Alta home for over 20 years and despite their best hopes, still, have a nearly impossible chance of owning the land on which their homes sit. As a result, many of these structures lack stewardship, the inability to own does not reasonably warrant investment and in turn, the homes exist in squalor.

The story of those in the upper mountain portions of Pamplona Alta is a different, yet only a slightly less dire narrative. For people who have built their homes on government-owned land, ownership is possible, however, only after a laundry list of nearly impossible requirements, given by the government of Peru, is achieved. Land ownership is possible to those who squatted on government land if they have:

- been there for longer than 5 years
- access to water
- access to electricity
- safe access to home
- community centers/parks within close proximity of their house

In practice, the government of Peru is asking the poorest and least advantaged people in their country to not only sustain life on a few dollars a day, but also build a road to create access for a privately priced water truck to get to their home, to pay to bring privately priced electricity to their home, to build community centers and to construct safe access to their homes from the bottom of the mountain.

The government task list is absolutely impossible on their own, and organizations like Solidaridad en Marcha have helped make home ownership a reality to some of the people of Pamplona Alta, however, this battle is still uphill, and full of rocks and mud.

The subsequent reality for the people of Pamplona Alta is that their children are frequently sick, their jobs are too far away, their under-education is inevitable, many families are broken and the lack of government support ensures the existence of a dark economy, thus extending their lack of access to upward mobility.

Deeper than these economic indicators of poverty, the people of Pamplona suffer from the poverty of dignity. Many members refer to themselves as “the forgotten ones,” election promises come and go without much change and through the confusing red tape and legalities, they still find themselves without access to basic human needs.

The Survey

Understanding the many facets of those in Pamplona Alta was integral in the way we built our survey, the questions we asked, and the way we asked them. We raised some questions that we heuristically had an intuitive idea of the answer, but needed to understand the severity. Yet, others we asked in order to gain insight into the tools we may have available to us within our solution set.

Lastly, we inquired of economic indicators, religious factors and family structure. All intended to paint a picture of the lives of those in Pamplona Alta and to possibly determine causality between the various characteristics.

Using this form, we were able to fully survey over 500 families and after extracting all personably identifiable data, we built a dataset that held great potential for a greater understanding of the lives of those in Pamplona Alta and the possible routes available to help them.

Número de la familia _____

La caja del amor
Un regalo de familia a familia

Solidaridad en Marcha

AAHH _____
Encargado _____
Celular _____

	NOMBRE	APELLIDOS	RELACIÓN	NOMBRE	SEXO (F/M)	EDAD
MADRE						
PADRE						

DIRECCIÓN Mz._____ Lt._____

Marca una 'x' según sea "Sí" o "No"

¿Tu teléfono tiene internet?	Sí	No
¿Puede la cisterna de agua llegar a tu casa?	Sí	No
¿Tienes una cuenta en el banco?	Sí	No
¿Vas al menos una vez al mes a la iglesia?	Sí	No
¿Sueles dejar a tus hijos solos en casa?	Sí	No

Responde solo con números

	Número
¿Cuántas personas viven en tu casa?	_____
¿Hace cuánto tiempo vives en esta casa?	_____
¿A qué edad tuviste tu primer hijo?	_____
¿Cuántas personas trabajan en tu casa?	_____
¿Cuánto tiempo te toma llegar a tu trabajo?	_____
¿Cuántos días de colegio pierden tus hijos al mes?	_____
¿Cuál es el ingreso mensual de tu familia?	_____
¿Cuántas personas en tu casa son bautizadas?	_____

The Data

The completed dataset built from the original survey contains 21 variables and 507 observations of which to explore, visualize and perform analysis on. The complete dataset can be found on my GitHub account, here. Also note, for binary variables, 1 is yes and 0 is no.

1. **fam_n** – *numeric factor* – Unique identifier for each family.
2. **internet** – *binary* – Does your phone have internet?
3. **agua** – *binary* – Can the water truck get to your house?
4. **banco** – *binary* – Do you have a bank account?
5. **iglesia** – *binary* – Do you go to church at least once a month?
6. **dejar_hijos** – *binary* – Do you leave your children home alone (when you go to work)?
7. **cuantas_personas** – *numeric* – How many people live in your house?
8. **tiempo_casa** – *numeric* – How long have you lived in your house?
9. **primer_hijo** – *numeric* – At what age did you have your first child?
10. **cuantas_trabajan** – *numeric* – How many people in your house work?
11. **tiempo_trabajan** – *numeric* – How long does it take to get to your job?
12. **pierden_colegio** – *numeric* – How many days a month do your children miss school?
13. **ingreso** – *numeric* – What is your monthly household income?
14. **bautizadas** – *numeric* – How many people in your family are baptized?
15. **direccion** – *character factor* – Name of neighborhood.
16. **padre** – *binary* – Does the father of the children live in the home?
17. **madre** – *binary* – Does the mother of the children live in the home?
18. **F** – *numeric* – Count of females in the home.
19. **M** – *numeric* – Count of males in the home.
20. **niños** – *numeric* – Count of children 18 and younger in the home.
21. **mayores** – *numeric* – Count of adults 65 and older in the home.

The Data Science Process

Months of conversations, meetings, reading and collaboration with community members came into building this survey. Qualitative analysis helped us produce a dataset that has the potential to perform multiple quantitative analyses that are relevant and informed; and it is from this point that we will now follow the flow of a data science analysis.

Reminder: I will at times use technical language, but I encourage you to keep reading through, because I will also accompany every scientific insight with an explanation in simple language, relevant to the question at hand.

We will begin by exploring the variables, the average values and basic correlations, visualizing how characteristics behave with one another. Following our exploratory phase, we will inspect cause and effect relationships between pairs of variables, as well as predict specific variables using all available data. I will use multiple techniques to perform this analysis of causality in hopes of providing variable importance in the prediction of key factors of the poor. The result will be a set of priorities for aid workers to pursue in the betterment of certain economic or societal indicators.

The succeeding analysis will be that of understanding natural segments that exist within the poorest of the poor. Again, using multiple techniques, I will attempt to determine the groups of people that exist within those surveyed. What commonalities do certain segments have? How can we target aid campaigns to help certain groups? These are a few of the many questions a segmentation analysis will help us answer.

Our quantitative and qualitative analyses will come to fruition in the final recommendation portion of this process. We will present questions and provide actionable insight into those questions, as determined by our analysis. We will build a road map for aid, a list of how we can help, who we can help and the logistical suggestions to do so. Our intent is to tie every insight to action and offer suggestions as to the best action available given what we have learned from the analysis. So, let’s get started!

Proportions of Binary Variables

We can see that a small proportion of those surveyed had bank accounts and internet access, while a large proportion went to church regularly and a majority had access to drinkable water though not overwhelmingly so. In terms of family dynamics, we see that only half of fathers are present in the home and most homes have a mother present.

Have Internet Access 11.06%	Have Access to Water 70.00%	Have Bank Accounts 3.89%
Regularly Attend Church Services 88.71%	Have Child Care 57.45%	Have a Father Present in Home 54.82 %

The initial pulse we get from the binary proportional averages is that we can affirm some of our pre-conceived ideas surrounding broken families and presence of a dark economy implying sparse routes to traditional credit sources.

Understand Average Values of Numeric Variables

Inspecting the numeric means of our survey results, we can begin to get a sense of the lives lead by the poorest of the poor in Pamplona Alta. We see that the average age of the mother when she had her first child is 21, which is young by developed world standards, but not nearly as young as our qualitative guess would have assumed; thus looking into the distribution of this variable could provide further insight.

Household Size 5 people	Time in Home 14 years	Mother's Age at First Child 21 years old	Work Commute 1 hour and 10 minutes
Working in House 1 to 2 people	School Missed per Month 2 days	Household Income per Month 649 Soles	Baptized in Household 3 people
Women in House 3	Men in House 2	Children in House 3	Elderly in House 0 to 1

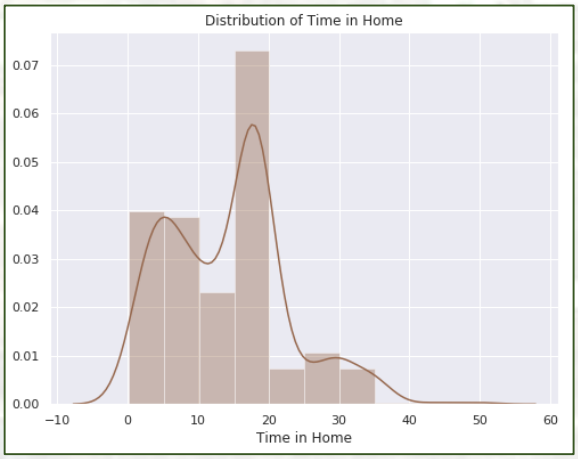
We see that the average time in the home is nearly triple the required 5-year requirement of the government for ownership, which is one good step forward. However, we also see that the average household of 5 is usually living off of the salary of a single person, and that average value is only 649 soles a month ($6.30 USD a day per household or $1.26 per person). Looking to the religious aspects of the survey, it is also interesting to note that on average only 2/5 members of households are baptized though a majority attend church services regularly.

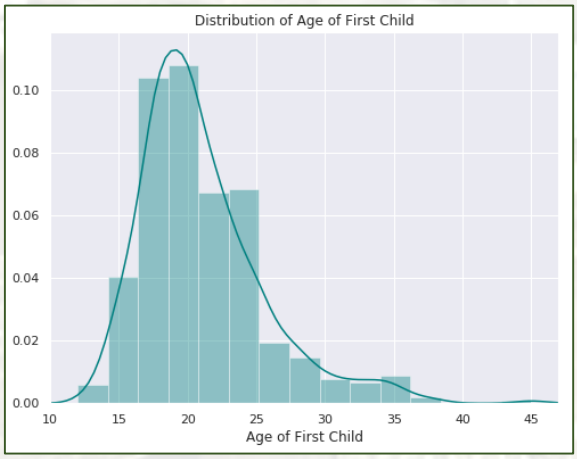
We do not want to make any extraneous assumptions from these average values, nevertheless, these initial figures have allowed us to paint a faint picture of the lives led by those surveyed and provided a question set for us to move forward in our analysis.

Variable Distributions

Average values can give us a glimpse into the character of a variable, however, the distribution can tell us even more. Given the above results I wanted to take the time to look into certain variables of intrigue. This process gives shape to the values within, as well as helps us spot outliers that may have significantly affected our larger groups at whole.

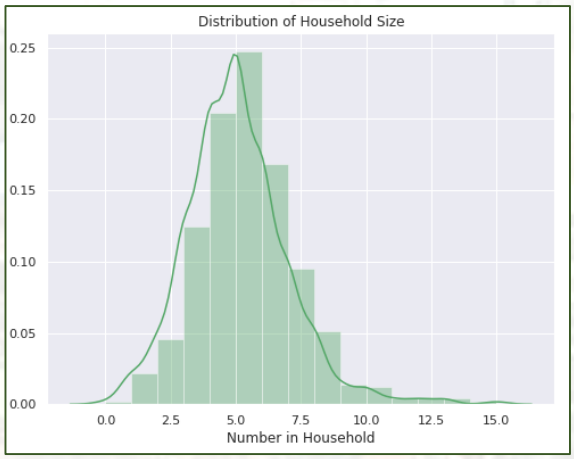
We can see that distribution of time-in-home is tri-modal, meaning that there are three time groupings in which people have been in their homes: many have been there less than 10 years, some more than 25, but most have been in their homes between 15-20 years. This is an encouraging data point when considering the government requirements for home-ownership.

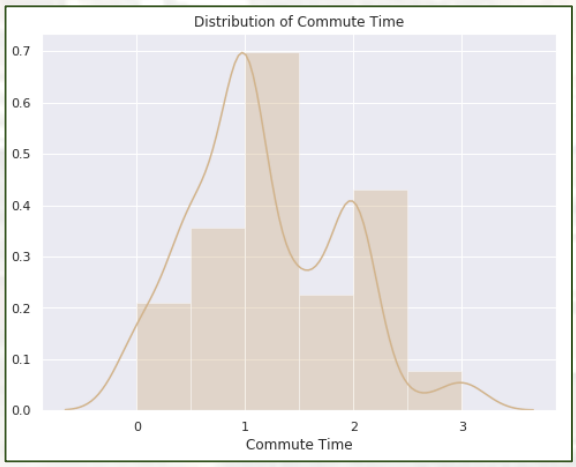




The second distribution, is that of the first child. The average value for this number was 21, which given our knowledge of the region, felt higher than we had experienced. The distribution confirms the fact that a majority of the women are having their first child at or before the age of 20. The average value was being pulled upwards due to the positive skew of the distribution. In Pamplona Alta, the reality of having children before the age of 17 is often dropping out of primary school, so to see fewer of these than we anticipated is mildly encouraging. However, the age of the first child in each household most certainly hinders further education beyond primary school for women.

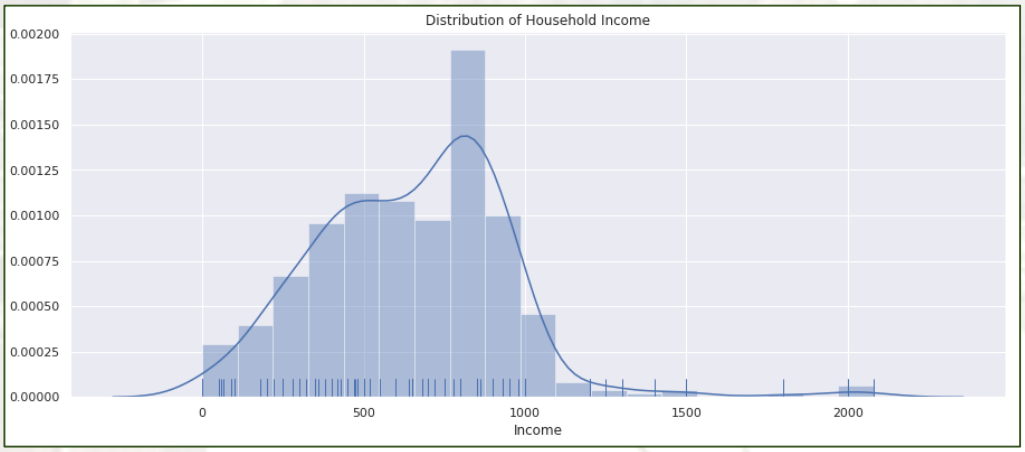
The third distribution is household size, the values are, more or less, normally distributed, with a majority of the households being between 4 and 6 people per household. We can see that there are some houses with as many as 15, but that number is small enough that those outliers should hold little weight to our predictive modeling further on.





Next, looking to commute time, we see a mostly bi-modal distribution, with a majority of people traveling just over an hour to work, and another traveling over two hours to work. This confirms the difficulty we've seen for people in Pamplona Alta to find employment near to home.

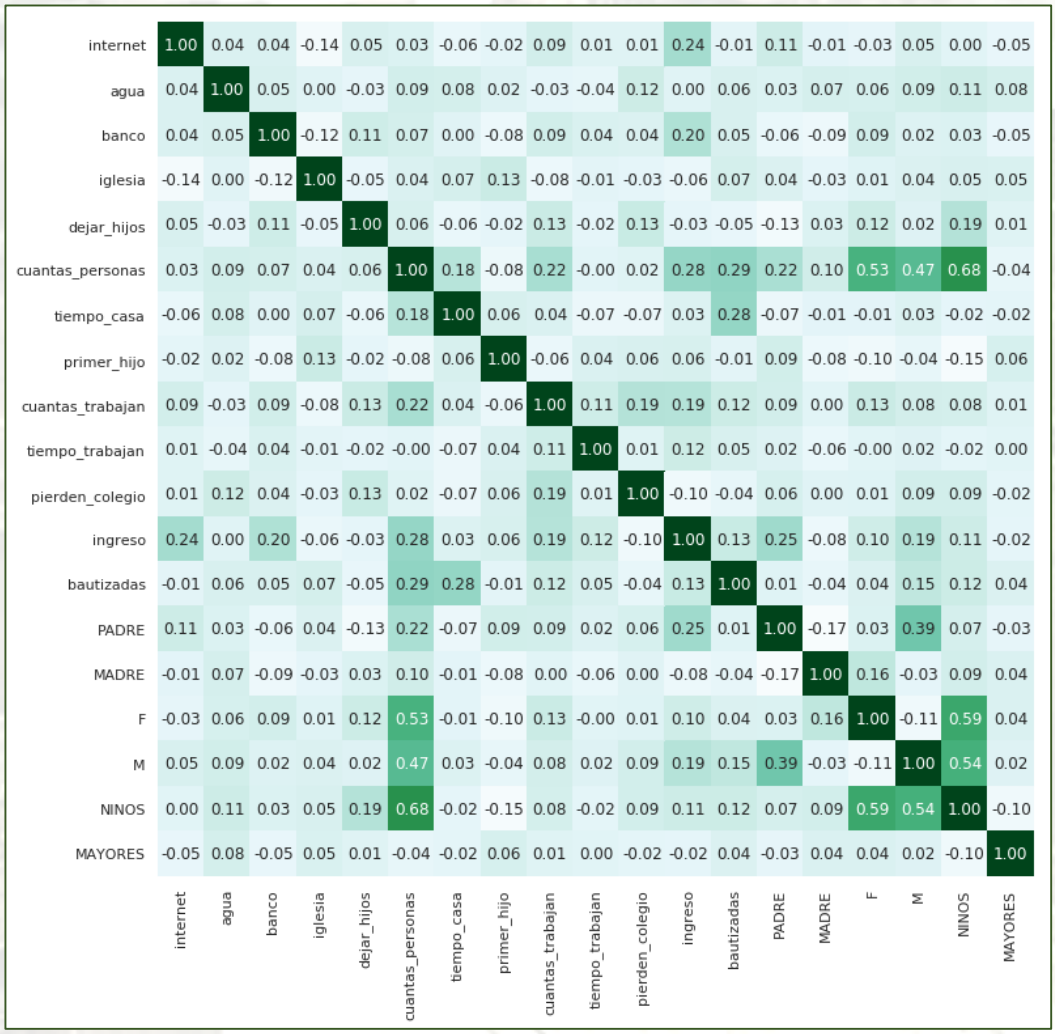
When attending to the distribution of household income below, we need to keep in mind that we removed two outliers early in the exploratory process. Nevertheless, we see that a majority of the families surveyed had a household income between 500/S and 1000/S Soles per month ($147.84 to $295.68 USD). Most estimates for income per capita in Peru are mostly north of 1500/S per month, therefore we can affirm that our survey truly touched the poorest of the poor.



Understand Relationships: Correlation

Beyond univariate analysis, the studying of only one variable, we want to start looking into how different factors behave in comparison to one another and to do so we will initially use correlation. Correlation is a mutual relationship, or connection between two or more characteristics/variable, therefore when comparing numeric relationships with correlation, the range is from -1 to +1.

If two variables are perfectly correlated, they will have a correlation of 1, while if they are perfectly opposite, they will have a correlation of -1. If there is no correlation, they will have a 0 value for correlation. In the chart below is every variable we have available for comparison, and matching values on the X axis with those on the Y, we can see the correlation value between those two values.



Initially, the correlation strengths are discouraging, we are seeing largely weak correlations between most of the variables, meaning that the predictive power of models we will build could be difficult to come by. Nevertheless, we see some relationships with more than weak positive and below we will dig into a few of these.

- The more time someone has been in their house, the higher the likelihood that they will be baptized. Thus, implying those who are established in a community are also established more in the local church, while newly established / less stable families have a more difficult time attaining access to the sacraments.
- If the father of the household is present there is not necessarily more workers in the house, but the household size and income are greater when he is present. What this could mean is that having a father present allows the family more money, thus the ability to support a larger household size.
- The relationship between internet and income is one of the stronger relationships, however the correlation does not mean causality. It is likely assumed that one needs money to pay for internet, however, it could be interesting to measure the causality of this relationship in order to see if access to internet also has the ability to propel the household income after attaining it. This is the same case for bank accounts, insofar as people with the most income have bank accounts, but with this analysis we do not know if bank accounts are causing higher income; for this, other analysis will be needed.
- We see that the younger the mother was when she had her first child, the more children she ended up having in the household, also those who became mothers earlier are less likely to be involved in the local church.
- The larger the household size, the longer time in the house. Suggesting that the larger the family the more established they are in the community and all the subsequent implied externalities.

It is important to note the strength of these hypotheses are only as strong as the underlying correlations behind them, however the exercise above gives us an opportunity to move forward and ask better questions.

Understand Relationships: Regression

The correlation gives us a high level understanding as to how two variables behave with one another, but the nature of that relationship is much deeper than correlation. Regression attempts to measure the relationship between a dependent variable and a predictor, or independent variable.

Simple regression essentially means, given characteristic ‘A’ (independent) how does characteristic ‘B’ (dependent) behave. The output of this analysis is often in a scatter plot with a trend-line to display the relationship. Other results that are important to pay attention to out of a regression analysis is the R-squared value, which helps us measure how much we can explain and how much we can’t.