Modeling Analysis

2023-04-29

Introduction

Given the background and information on refugee migration into Moldova from Ukraine, a questions arises: where will these refugees go in the near and far future? The respondent's intentions of their future migration plans is the primary focus of this analysis. We aim to address the question:

What factors are associated with a refugee's intention to either return to or leave Ukraine?

Given this intention, our response variable of interest is whether the individual plans to return to Ukraine or not. If the respondent intends to return to Ukraine (whether to their home oblast or other), the response was recorded as a "yes" (1); if the respondent plans on staying in Moldova or moving to a different country, the response was recorded as "no" (0). Using this criteria, 553 respondents do not plan to return to Ukraine, 73 do plan to return to Ukraine, and 78 either didn't respond or did not know. For the sake of this analysis, we will limit this scope to only respondents that responded yes or no. After accounting for missingness in other attributes associated with this study, there are 547 unique and complete observations in which we will explore. Eaxh observation represents the interview/survey of a unique Ukrainian refugee.

To provide further detail on the variables of interest, exploratory data analysis can be seen below.

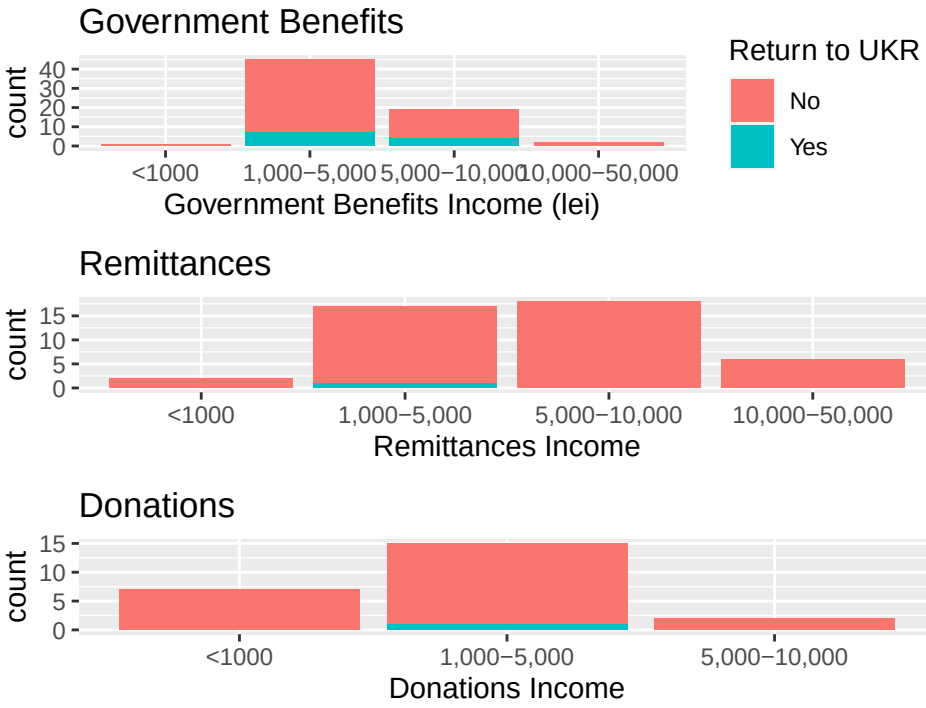
Exploratory Data Analysis

To begin, we will observe the income distributions by return intentions for respondents. As seen in Figure 1, although there are very few data entries that were not N/A for business income values, we can see that mid-level incomes correlated with a great number of respondents noting that they intended to return to Ukraine. Such finding may suggest that moderately successful business owners may want to continue way of life where as wealthier, more secure individuals feel more comfortable leaving. Similar to business income, there are very few data for salaried income. In fact, there are only 2 respondents that have salary data and repsponded that they plan to return to Ukraine. These findings motivate us to be cautious if including these variables in the model as there is not much data.

Figure 2 covers income categories that had more observations, such as income from government benefits, remittances, and donations, however, once again, there is little variation to be seen within income levels by return intention. Due to the low number of observations for each income category, in our modeling process, we will combine all income counts into a "total income" variable.

Figure 1: Income Distributions by Return Decision

Figure 2: Income Distributions by Return Decision



An additional part of our analysis that we are concerned with is familial characteristics of respondents. Figure 3 below shows that respondents intending to return to Ukraine had smaller family sizes with them than that respondent's that do not intend to return. The median family size for those that intend to return is 2 whereas the median family size for those that do not intend to return is 3. This may suggest that having more children is associated with the intention not to return to Ukraine. However, respondent's may also not have taken their entire family.

Figure 3: Respondent's Family Size by Return Intention

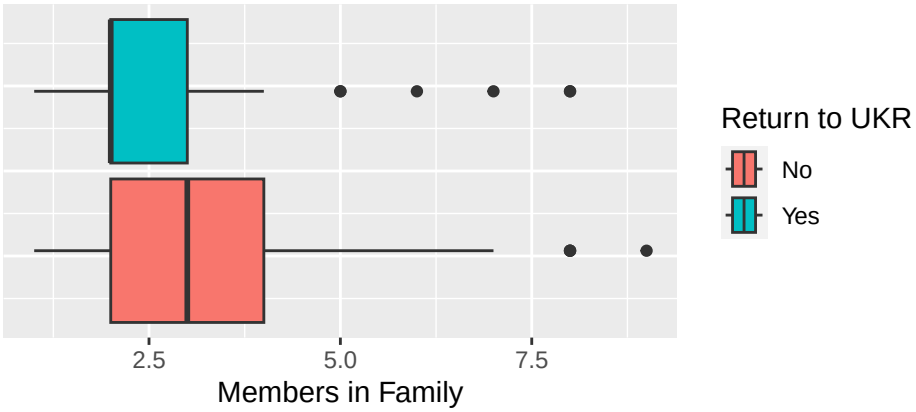


Figure 4 below shows the distribution of the presence of a family member 6 months old or younger and shows the respective return intention. As seen below, less than 3.0% of respondents had an infant in their family. With such low presence, a lack of meaningful variation will keep us from including this variable in our modeling process, but it is still an interesting finding worth noting.

Figure 4: Infant in Respondent's Family by Return Decis

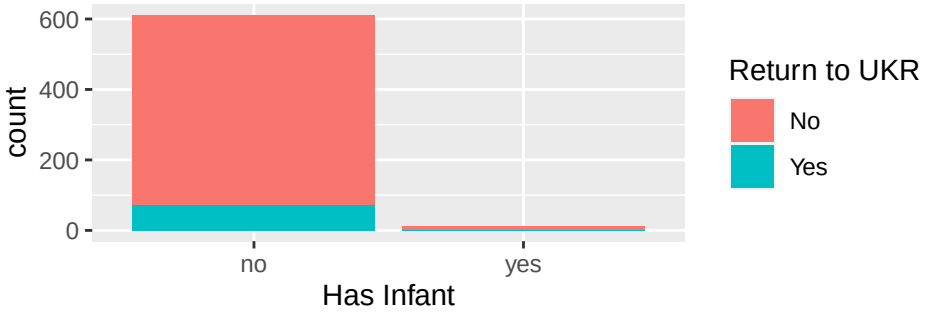


Figure 5 below shows that respondents that do not intend to return to Ukraine have a smaller range in distribution of the respondent's age. For those that do not intend to return the minimum age is lower and the maximum age is higher compared to the respective metrics of those that do intend to return. However, the median age of those not intending to return is lower than those who do intend to return.

Figure 5: Respondent's Age by Return Intention

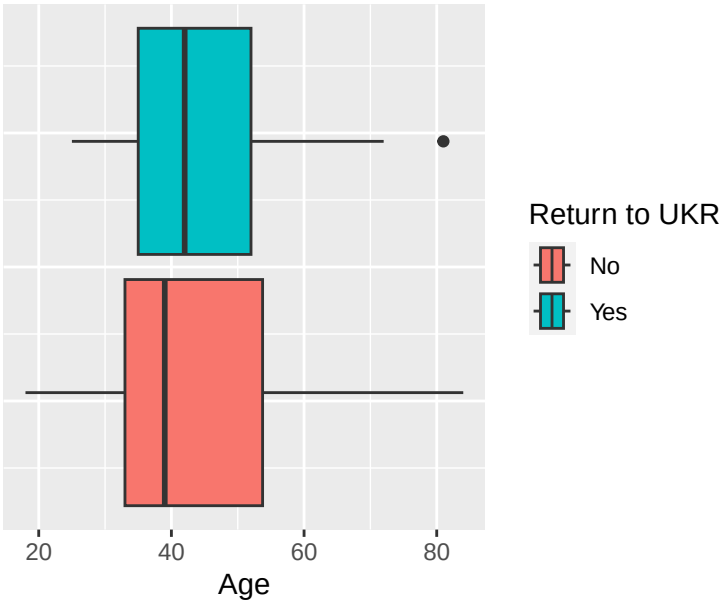


Figure 6 evidences another variable in which the highly saturated response negates it from our modeling analysis, but the findings are interesting nonetheless. This figure represents whether every member of ther respondent's family has a valid ID/passport. More than 90% of the respondents noted 'yes', and a higher proportion out of the 32 respondents that responded 'no' intend to return to Ukraine. This may suggest that a lack of ability to travel elsewhere is affecting their return plans.

Figure 6: Official ID available for entire family

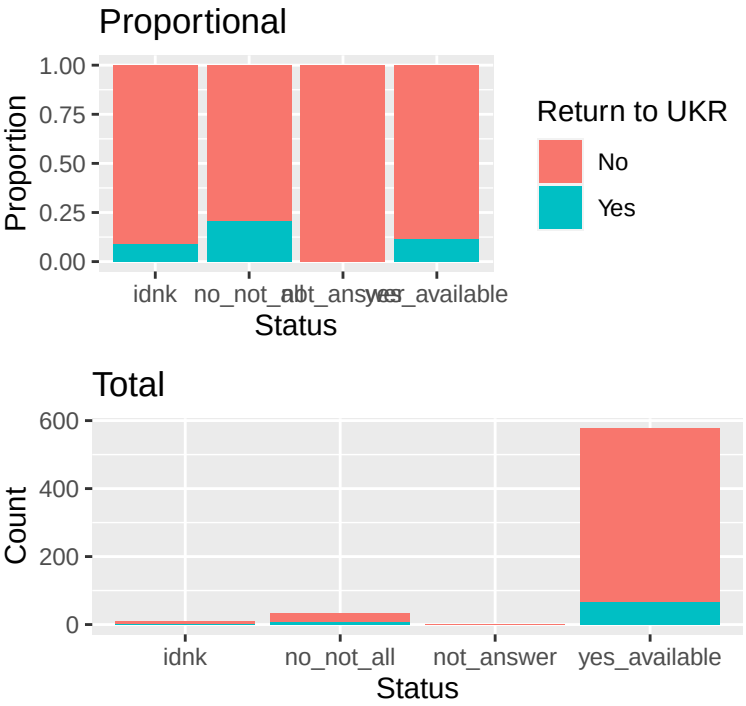


Figure 7 below shows the highest education level of the respondent's household head by return intention proportion. Most respondents had a household head that had completed higher education. However, there is little significant vairation in proportion of return intention by educaiton type.

Figure 7: Education level by Return Decision

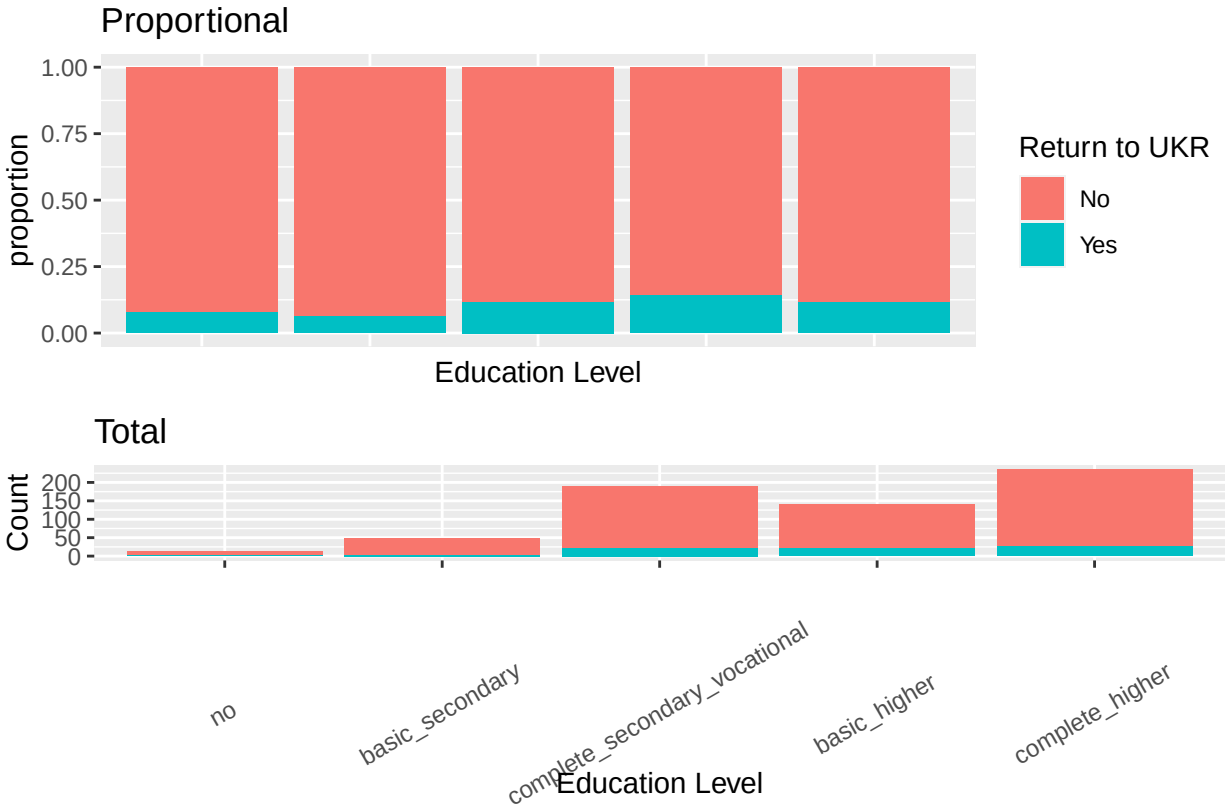
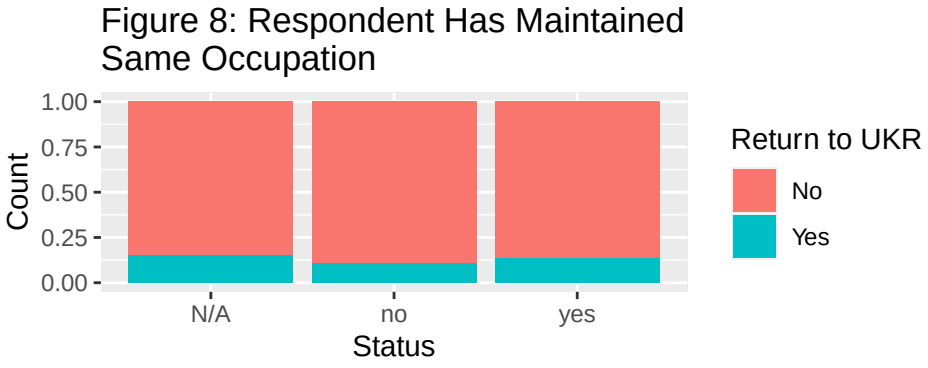
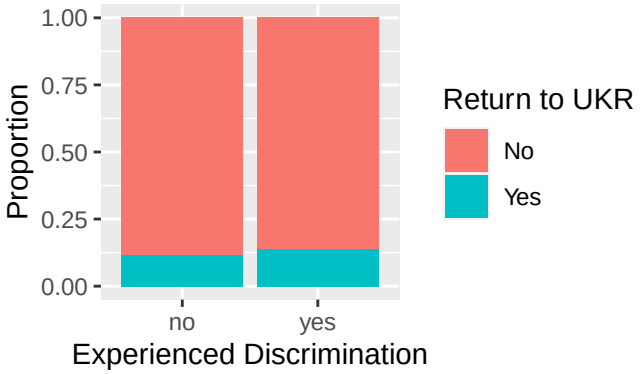


Figure 8 represents the occupation changes in Ukrainian refugee respondents. While most of the respondents, as expected, are not currently in the same occupation since arriving to Moldova, a greater proportion of those that are intend to return to Ukraine This may suggest some sort of stability that draws them back to their home country.



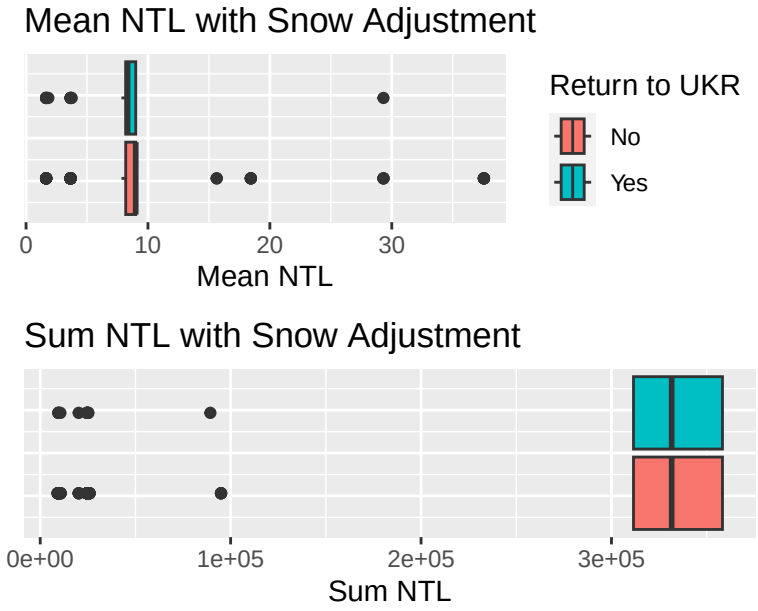
As seen in Figure 9 there is little difference in discrimination experience across respondents who intend versus do not intend to return to Ukraine. And there were few respondents who had faced any discrimination ($<$ 10%).

Figure 9: Experienced Discrimination
]by Return Decision



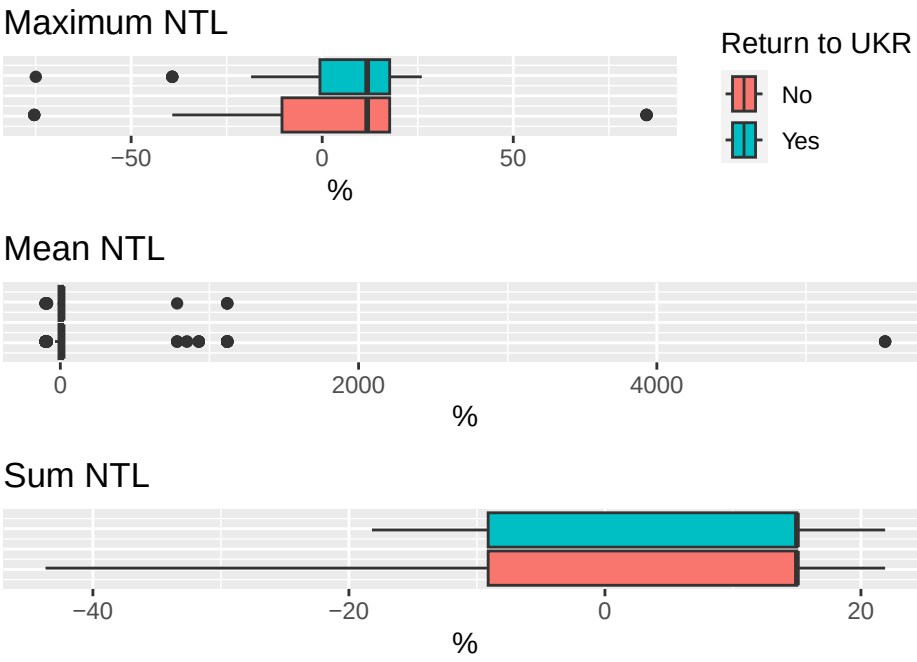
Another area this analysis explores is the Night Light activity of the areas in Moldova of which the respondents are in. For this analysis, these areas are Bender, Balti, Chisinau, and Transnistria. Specifically, we looked at night light activity during the month of the respondent's arrival within the respective area in Moldova. Figure 10 demonstrates below that there was virtually no difference in the NTL sum or mean of the respondent's arrival month in the location of Moldova between respondents intending to return to versus leave Ukraine. This may be a biproduct of the fact that this data only spans a few months. Perhaps, if we had data over a year or more we would see more variation.

Figure 10: NTL data Compared to Return Intention



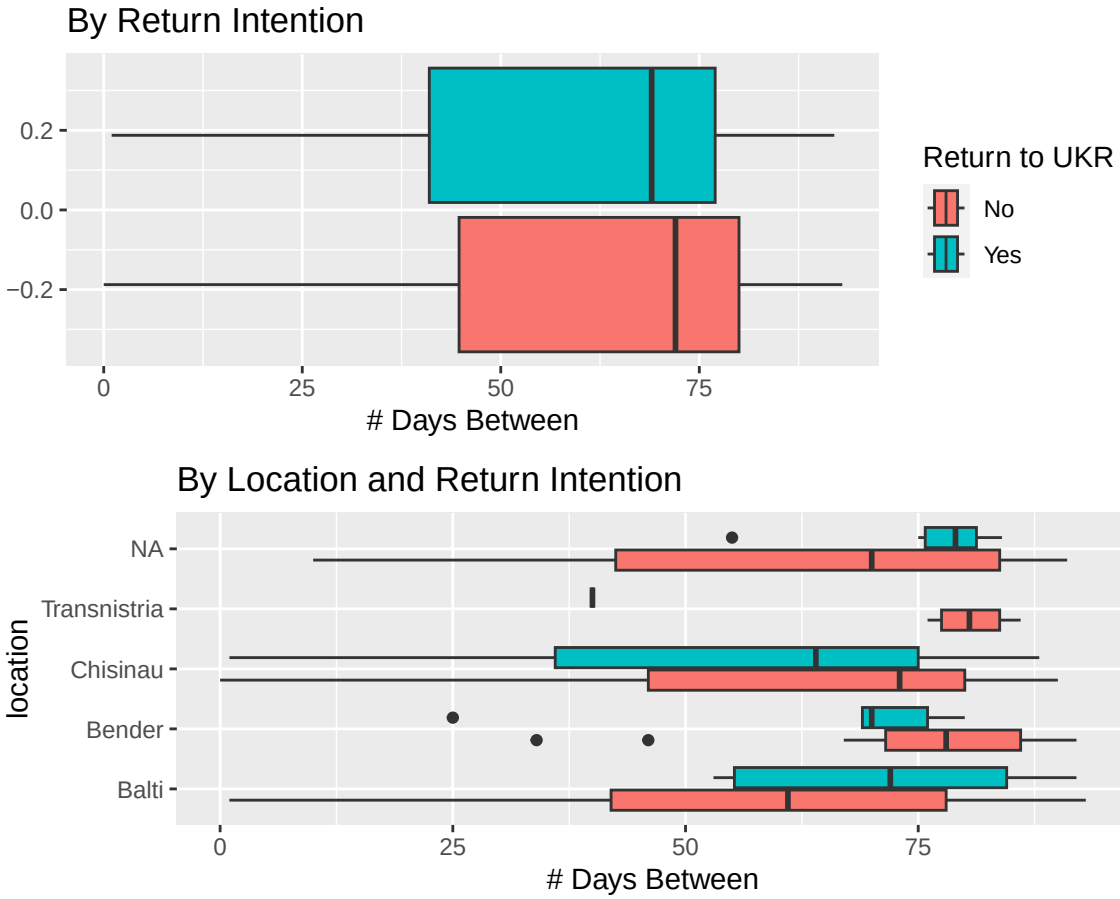
Continuing our analysis of NTL activity, Figure 11 below represents the percent changes in different NTL measurements of the month of the respondent's arrival to the month before their arrival in a respective area. There is little variation to be seen between return intentions of respondents. Again, with more data spread across more than just several months, notable variations may be present.

Figure 11: NTL Percent Changes from
month prior arrival to arrival month



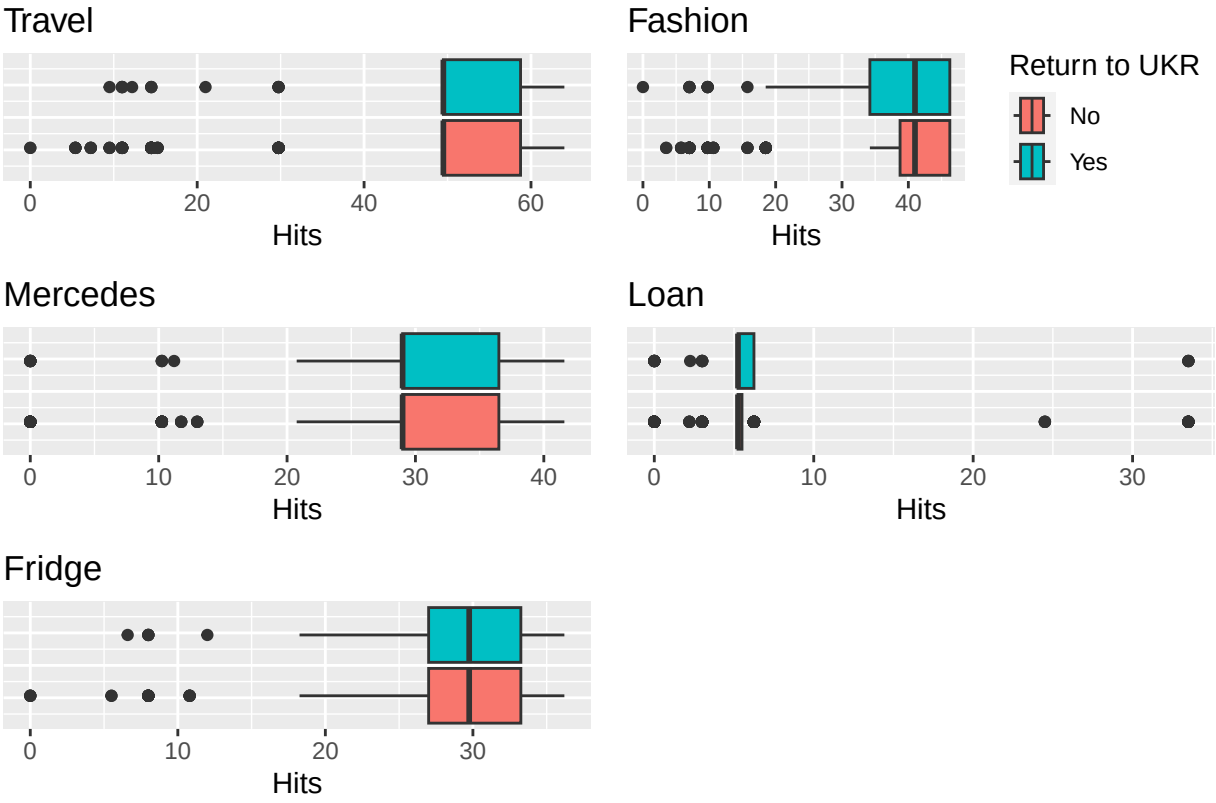
The first figure Figure 12 below suggests that there is no notable difference in the number of days between arrival and the interview in regards to intention on returning to Ukraine. Although, the median of those with no intention of returning is several days higher than those with the intention of returning. However, there is noticeable variation when the figures are separated by location in Moldova. For respondents where there was no recorded location or in Balti, the distribution of days between arrival and interview for respondents not intending to return to Ukraine is less compared to respondents intending to return. The reverse is seen for Transnistria, Chisinau, and Bender.

Figure 12: Days Between Respondent's Arrival and Interview



Google Trends data was another component of this analysis we were interested in. Figure 13 below represents the Google trends number of hits for a given topic in the month of a respondent's arrival in the location of the respondent within Moldova. Because our data is monthly,and the respondent's arrival dates are only distributed across only several months, there is little variation seen between the respondent's that intend to return versus leave Ukraine. Additionally, Google trends also picks up the searches of the existing/indigenous popualtion. With more data, or with the availability of daily Google Trends hits, we may see more variation between intention groups.

Figure 13: Google Trends Data Showing Some Variation



Modeling

For our modeling, we will fit a logistic regression model with response as the binary vairable indicating whether a respondent intends to return to Ukraine or not. Our data will be a combination of the UNHRC data and the Google Trends and NTL data matched based on arrival month and location of the repsondent.

Variables of interest

The following variables were of particular interest when considering the composition of our model

- Location - The current location of the respondent in Moldova (Chisinau, Bender, Balti, or Transnistria)
- Language preferences - binary variables indicating if the speaker prefers Romanian, Russian, or Ukrainian
- The age of the speaker
- Household accomodations - where the respondent is currently staying (hotel, with relatives, accredited refugee accomodation center (RAC), etc.)
- Origin Ukrainian oblast
- Binary representing whether, after the start of the war, did the respondent have to move to another place in Ukraine before coming to Moldova.
- Conditions the respondent would need to change in Ukraine before you decided to return?

- Number of people in the respondent's family
- Binary variable indicating if, since arriving to Moldova, has the respondent applied to enroll their child in school/kindergarten in Moldova?
- Number of people in the Moldovan household the respondent is staying in
- Marital status of the household host
- Binary representing if any member in the respondent's family (Ukrainian) is six months of age or younger?
- Binary variables indicating if the following items are limited: adult clothing, basic hygiene, beds, heaters, children clothing, diapers, or menstrual materials.
- Highest level of education attained by the head of respondent's family
- Whether or not the respondent needs to learn / improve a language to integrate in Moldova work market
- Occupation before the war began
- Sector of occupation before the war began
- If the respondent is in the same occupation since arriving in Moldova
- The respondents current occupation
- If the respondent needed to sell household assets/goods (radio/furniture/TV...)
- If the respondent needed to withdraw children from school
- If the respondent needed to sell their house/land
- Total income from salary work, daily labor, personal business, savings/pension, government benefits, remittances, support from friends and family, donations, and humanitarian assistance
- Total expense amount on food, rent, water, non-food household items (hygiene, lightbulbs, etc), utilities, fuel (for cooking, vehicles, etc.), transportation, and communication
- Indicator whether or not every person in the respondent's family has an ID document (national ID and/or passport)?
- Indicator whether or not the respondent or anyone in their family experienced what discriminatory treatment since arriving to Moldova?
- The month and the year of the respondent's arrival into Moldova
- The month and year which the interview took place
- Number of days between the arrival and interview
- The following are "hits" from google trends. Each trend was measured from 2018 to the end of 2022. The "hits" represent a normalized distribution of "hits" of a certain topic. These numbers don't represent absolute search volume numbers, because the data is normalized and presented on a scale from 0-100, where each point on the graph is divided by the highest point, or 100. A lower number of "hits" means that a search term's relative popularity is decreasing—not necessarily that the total number of searches for that term is decreasing, but that its popularity compared to other searches is shrinking. This "hits" value was found for both the month of the respondent's arrival (taken as an average of weeks within the month) as well a month before the respondent's arrival. Using this criteria, the following topics are included:

  ** Travel ** Holiday ** Labor ** Unemployment ** Loan ** Washing Machine ** Fridge ** BMW ** Mercedes ** Fashion

- We will also be using night light data (snow adjusted) from the month of the arrival of the respondent. The following metrics will be used: ** Maximum ** Mean ** Sum ** Standard Deviation
- A percent change in these metrics from the month prior to arrival will also be used.
- In addition to this nightlight data, percent changes of the these metrics from the previous month will be used

Model Formula

This is the full mathematical formulation of the full logistic model produced in this study.

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0 + \beta_1 x_{\text{Prefer Ukrainian}i} + \beta_2 x_{\text{Prefer Russian}i} +$$

$$...\beta_3 x_{\text{Total Expenses btwn 1,500-3,000 (vs. <1000)}i} + \beta_9 x_{\text{Total Income btwn 1,000-5,000 (vs. <1000) lei}i} + \cdots$$

$$... + \beta_{15} x_{\text{Travel}i} + +\beta_{16} x_{\text{Fashion},i} + ... + \beta_{20} x_{\text{Respondent Age}i} + \beta_{21} x_{\text{Respondent's family size}i} +$$

$$... + \beta_{28} x_{\text{Beds Limited}i} + ... + \beta_{48} x_{\text{Fridge hits (month prior to arrival)}i}$$

π_i is the "success probability" for call i. This means the probability that a respondent intends to return $\frac{\pi_i}{1-\pi_i}$ is the odds ratio, or the probability of intending to return over the probability of not. The index of observation i corresponds to each return intention. β is a vector of logistic model coefficients in the order of "Model Output" below. For example, β_0 is the coefficient of the intercept. Similarly, β_3 is the coefficient of the `Total Expenses btwn 1,500-3,000 (vs. <1000)` (x_3) variable and so on.

Model Output

Table 1: All Variables (Full Model)

term	estimate	Confidence (2.5%)	Confidence (97.5%)	P-Value	Significant
Intercept	5.066e-309	0.000e+00	Inf	0.983	No
Prefer Ukrainian	2.603e-01	4.412e-02	1.535e+00	0.137	No
Prefer Russian	3.313e-01	3.127e-02	3.509e+00	0.359	No
Total Expenses btwn 1,500-3,000 (vs. <1000) lei	1.455e+00	2.416e-01	8.763e+00	0.682	No
Total Expenses btwn 10,000-15000 (vs. <1000) lei	1.053e+00	1.315e-01	8.442e+00	0.961	No
Total Expenses btwn 3,000-7,500 (vs. <1000) lei	1.396e+00	2.568e-01	7.586e+00	0.699	No
Total Expenses btwn 500-1,500 l(vs. <1000) ei	1.341e+00	1.718e-01	1.047e+01	0.779	No
Total Expenses btwn 7,500-10,000 l(vs. <1000) ei	8.664e-01	1.101e-01	6.819e+00	0.892	No
Total Expenses unknown (vs. <1000) lei	1.255e+00	2.282e-01	6.906e+00	0.794	No

term	estimate	Confidence (2.5%)	Confidence (97.5%)	P-Value	Significant
Total Income btwn 1,000-5,000 (vs. <1000) lei	5.230e-01	8.183e-02	3.343e+00	0.493	No
Total Income btwn 5,000-10,000 (vs. <1000) lei	4.400e-01	6.121e-02	3.163e+00	0.415	No
Total Income btwn 10,000-50,000 (vs. <1000) lei	1.051e+00	1.527e-01	7.232e+00	0.960	No
Total Income btwn 50,000-100,000 (vs. <1000) lei	4.397e-09	0.000e+00	Inf	0.999	No
Total Income > 100,000 (vs. <1000)	4.733e-09	0.000e+00	Inf	0.999	No
Total Income unknown (vs. <1000) lei	7.531e-01	1.347e-01	4.212e+00	0.747	No
Travel hits (arrival month)	8.259e+12	0.000e+00	Inf	0.983	No
Fashion hits (arrival month)	1.255e+16	0.000e+00	Inf	0.983	No
Mercedes hits (arrival month)	1.113e+06	0.000e+00	Inf	0.983	No
Loan hits (arrival month)	3.542e+08	0.000e+00	Inf	0.983	No
Fridge hits (arrival month)	3.705e+01	7.909e-138	1.736e+140	0.982	No
Respondent Age	1.004e+00	9.767e-01	1.031e+00	0.793	No
Respondent's family size	1.062e+00	8.605e-01	1.310e+00	0.576	No
% Change in Maximum NTL from one month prior arrival - arrival	4.266e+03	0.000e+00	Inf	0.983	No
Highest Education(HH): Basic Secondary (vs. none)	5.352e-01	3.394e-02	8.439e+00	0.657	No
Highest Education(HH): Complete Secondary Voc. (vs. none)	6.676e-01	5.523e-02	8.070e+00	0.751	No
Highest Education(HH): Basic Higher (vs. none)	9.528e-01	7.930e-02	1.145e+01	0.970	No
Highest Education(HH): Complete Higher (vs. none)	7.500e-01	6.091e-02	9.235e+00	0.822	No
Adult Clothing Limited	1.672e+00	6.186e-01	4.516e+00	0.311	No
Beds Limited	1.573e+00	4.518e-01	5.476e+00	0.477	No
Children Clothing Limited	8.789e-01	2.860e-01	2.701e+00	0.822	No
Diapers Limited	1.917e-09	0.000e+00	Inf	0.998	No
Heating Limited	4.227e-01	7.885e-02	2.266e+00	0.315	No
Cooking Items Limited	1.022e-01	1.276e-02	8.180e-01	0.032	Yes
Menstrual Material Limited	9.734e-01	3.047e-01	3.109e+00	0.964	No
Basic Hygiene Items Limited	2.729e-01	7.889e-02	9.442e-01	0.040	Yes
Nothing Limited	5.622e-01	2.328e-01	1.358e+00	0.200	No
Occupation Change NA (vs. changed)	1.153e+00	4.682e-01	2.839e+00	0.757	No
Occupation Remained Same (vs. changed)	1.211e+00	4.886e-01	3.002e+00	0.679	No
Current Occupation: Caregiver (vs. not working)	4.389e-01	8.791e-02	2.192e+00	0.316	No
Current Occupation: Formal Work (vs. not working)	5.104e-09	0.000e+00	Inf	0.999	No

term	estimate	Confidence (2.5%)	Confidence (97.5%)	P-Value	Significant
Current Occupation: Informal Work (vs. not working)	9.018e-01	7.525e-02	1.081e+01	0.935	No
Current Occupation: Retired (vs. not working)	8.440e-01	2.672e-01	2.665e+00	0.772	No
Current Occupation: Student (vs. not working)	3.523e-09	0.000e+00	Inf	0.999	No
Current Occupation: Volunteer (vs. not working)	4.318e-09	0.000e+00	Inf	0.999	No
Travel hits (month prior to arrival)	2.285e-11	0.000e+00	Inf	0.983	No
Fashion hits (month prior to arrival)	4.387e-30	0.000e+00	Inf	0.983	No
Mercedes hits (month prior to arrival)	3.068e-04	0.000e+00	Inf	0.984	No
Loan hits (month prior to arrival)	1.819e+14	0.000e+00	Inf	0.983	No
Fridge hits (month prior to arrival)	1.051e+16	0.000e+00	Inf	0.983	No

Result Discussion

As seen in the model table, the results were highly disappointing. For multiple estimates, the 95% CI for a logistic regression coefficient includes both 0 and infinity. This means that the coefficient is not significantly different from 0, and the odds ratio for a one-unit change in the predictor variable could be any value between 0 and infinity. This can occur when there is a lack of statistical power to detect a significant effect or when the predictor variable is not linearly related to the outcome variable.

We believe that because the time span for this data is relatively short (only a couple months) and there is a lot of missingness in the responses among an already relatively small dataset, there was a severe lack in variation amongst the predictors, thus resulting in an under-performing model. However, this methodology begs to be further built upon in later research. Collecting similar data as seen here and expanding upon it by ensuring more complete responses and over a larger period of time may allow for more meaningful future analyses to assess correlation between return intentions and other factors.

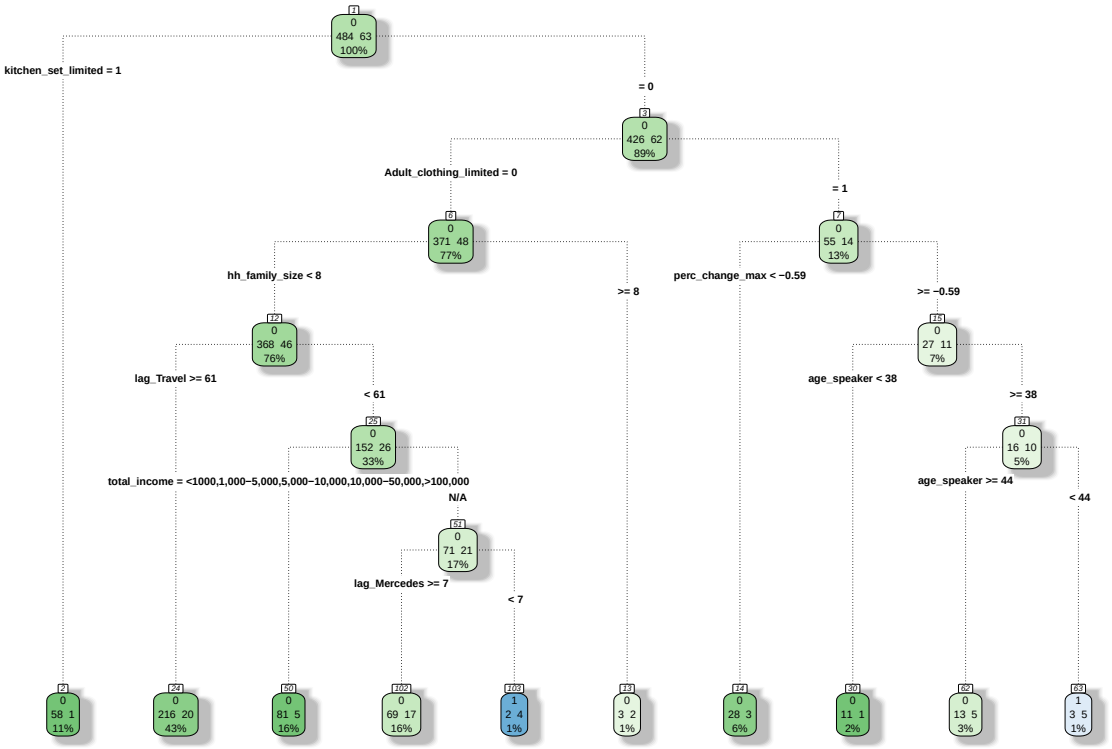
In our model, there were 2 variables deemed significant at 0.1 alpha level. The interpretations are provided below:

- **Basic Hygiene Items Limited** - Holding the other variables constant, when basic hygiene items are limited (versus not), the predicted odds of a respondent intending to return vs not are multiplied by a factor of 0.272.
- **Cooking Items Limited** - Holding the other variables constant, when cooking items are limited (versus not), the predicted odds of a respondent intending to return vs not are multiplied by a factor of 0.102.

Both these "factors" are less than 1, meaning the odds of intending to return to Ukrain are essentially decreased rather than increased.

Decision Tree

With no clear take-aways from our full logistic model, we will also try using decision trees to better visualize the relationship between return intentions and our predictors. The decision tree was constructed by originally including all variables in our previous logistic model (can be seen above).



Decision Tree Results Discussion

The overall accuracy of our model was **89.21%**, which performs better than random choice (50%). From our results, we can see that kitchen_set_limited was selected as the first variable to split the data into two subgroups. `kitchen_set_limited` is a binary variable representing if the respondent faced limited kitchenware or cooking items. As it is the first variable that splits the data, it is the one that has the highest predictive power. This finding is mirrored in our original full logistic model as it was one of only two variables to be significant. In contrast to our full model above, the binary variable representing if basic hygiene items are limited was not included in this decision tree. Instead, our variable with the second highest predictive power is the binary variable representing if adult clothing is limited. Other variables included in this resulting decision tree are as follows:

- `hh_family_size`: House hold family size - the number of people in the respondent's family that the respondent is with. This variable had an initial cutoff at above 8 members
- `perc_change_max`: The percent change in maximum snow-adjusted nightlight values from the month previous to arrival to arrival month. This variable had an initial cutoff at below 0.59 %.
- `lag_Travel`: The Google trends normalized number of hits from beginning 2018 through 2022 for the topic "Travel" for the month prior to the respondent's arrival. This variable had an initial cutoff at or above 61 hits.
- `age_speaker`: The age of the respondent. This variable had an initial cutoff at or above 38, and then within the older group (older than 37) a second cutoff at or above 44.
- `total_income`: The total income of the respondent. This variable had an initial cutoff at a "N/A" or not "N/A" level.
- `lag_Mercedes`: The Google trends normalized number of hits from beginning 2018 through 2022 for the topic "Mercedes" for the month prior to the respondent's arrival. This variable had an initial cutoff at or above 7 hits.