

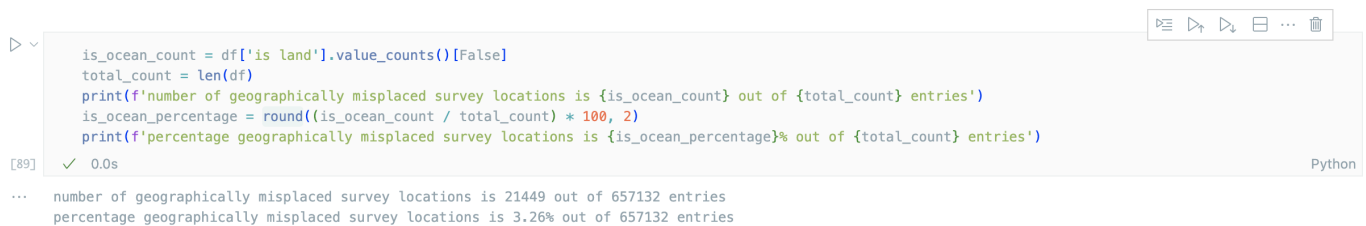
Cases to Explore

Poverty Stoplight Platform Database Analysis

Guidelines: Below are the types of issues that were identified in the poverty stoplight platform database. For each issue, I provide 1) a description of the issue as it was provided to me by my supervisor, Alexis Risaldi, and then 2) how I identified the issues in the dataset, and 3) how I answered the analysis questions, along with screenshots of the results for the three database excel files I had (attached in the zip file). Additionally, all the results and analysis that you see below are exported into new excel files (attached in the zip file) that you are welcome to look over. You can generate similar exported excel files by replacing the input excel files and matching the new excel file names into the Python script provided. Lastly, when using the Python script provided, please make sure to download any necessary libraries on your local device, and if you don't have the right setup to run Jupyter Notebooks locally on your device, you can upload the notebook into a Google Colab file, and then you'll be able to run the script remotely. Feel free to reach out to me, Aya Hajjeh, at ayabaselhajjeh@gmail.com for any troubleshooting assistance needed.

1. Surveys that are geographically misplaced

To identify (latitude, longitude) value pairs that are geographically misplaced, I used the library global-land-mask (credit to Karin, Todd. Global Land Mask. October 5, 2020. <https://doi.org/10.5281/zenodo.4066722>) to flag the pair values that are in the ocean and not on land.



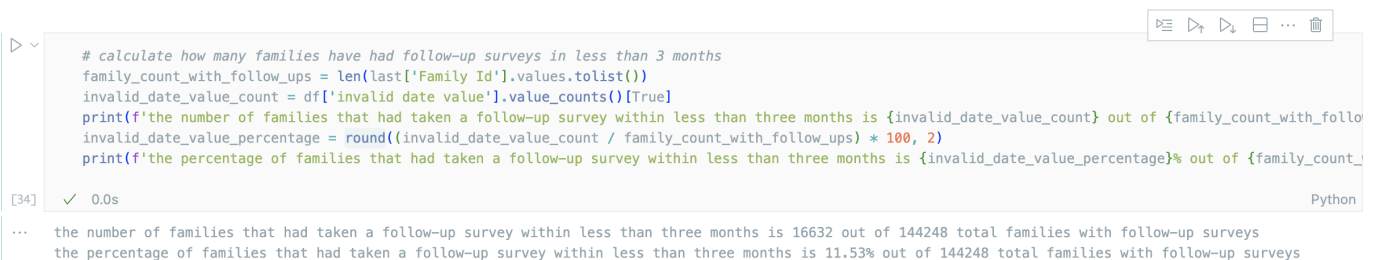
```
is_ocean_count = df['is_land'].value_counts()[False]
total_count = len(df)
print(f'number of geographically misplaced survey locations is {is_ocean_count} out of {total_count} entries')
is_ocean_percentage = round((is_ocean_count / total_count) * 100, 2)
print(f'percentage geographically misplaced survey locations is {is_ocean_percentage}% out of {total_count} entries')
```

[89] ✓ 0.0s Python

... number of geographically misplaced survey locations is 21449 out of 657132 entries
percentage geographically misplaced survey locations is 3.26% out of 657132 entries

2. Surveys that were taken in less than 3 months for the same person

Over the months, Poverty Stoplight team have found follow-up surveys with differences in minutes or days. Follow-up surveys should have a minimum difference of 3 months to analyze if there were any changes in the indicators. Therefore, the Python script I created flag any survey follow-ups that were taken in less than 3 months for the same person.



```
# calculate how many families have had follow-up surveys in less than 3 months
family_count_with_follow_ups = len(last['Family Id'].values.tolist())
invalid_date_value_count = df['invalid date value'].value_counts()[True]
print(f'the number of families that had taken a follow-up survey within less than three months is {invalid_date_value_count} out of {family_count_with_follow_ups}')
invalid_date_value_percentage = round((invalid_date_value_count / family_count_with_follow_ups) * 100, 2)
print(f'the percentage of families that had taken a follow-up survey within less than three months is {invalid_date_value_percentage}% out of {family_count_with_follow_ups}')
```

[34] ✓ 0.0s Python

... the number of families that had taken a follow-up survey within less than three months is 16632 out of 144248 total families with follow-up surveys
the percentage of families that had taken a follow-up survey within less than three months is 11.53% out of 144248 total families with follow-up surveys

3. Surveys that take less than 20 minutes to complete

The average time to complete the surveys is 20 minutes, which typically includes around 25 socioeconomic questions and an average of 50 indicators. To identify the surveys where this isn't the case, the Python script flags the surveys that were taken in less than 20 minutes.

```
total_time_less_than_20_count = df['invalid total time taken to fill out survey'].value_counts()[True]
total_count = len(df['invalid total time taken to fill out survey'].values.tolist())
print(f'number of surveys with less than 20 minutes total time taken to fill out the survey is {total_time_less_than_20_count} out of {total_count} total en
percentage_time_less_than_20_count = round((total_time_less_than_20_count / total_count) * 100, 2)
print(f'percentage of surveys with invalid less than 20 minutes total time taken to fill out the survey is {percentage_time_less_than_20_count}% out of {tot
```

[96] ✓ 0.0s Python

... number of surveys with less than 20 minutes total time taken to fill out the survey is 78 out of 657132 total entries
percentage of surveys with invalid less than 20 minutes total time taken to fill out the survey is 0.01% out of 657132 total entries

4. Surveys that are answered by children under 14

In the European Union, minors' personal data protection is regulated by the General Data Protection Regulation (GDPR). According to the GDPR, parental or legal guardian consent is required to process the personal data of minors under 14 years old. I identify all the surveys from minors under 14 years old and flag them in an exported excel file so we can take appropriate action.

```
age_count_under_14 = df['invalid age value'].value_counts()[True]
age_count_total = len(df['invalid age value'].values.tolist())
print(f'number of age participants under 14 is {age_count_under_14} out of {age_count_total} total entries')
age_percentage_under_14 = round((age_count_under_14 / age_count_total) * 100, 2)
print(f'percentage of age participants under 14 is {age_percentage_under_14}% out of {age_count_total} total entries')
```

✓ 0.0s Python

number of age participants under 14 is 39855 out of 657132 total entries
percentage of age participants under 14 is 6.06% out of 657132 total entries

5. Invalid Snapshot Values

Some data entries have invalid snapshot values and that could for either one of the following reasons: a missing snapshot value, a duplicated snapshot value, or a missing family id. I identified such invalid values, flagged them in an exported excel file, along with the reasons on why each survey had been flagged.

```
invalid_snapshot_value_count = df['invalid snapshot value'].value_counts()[True]
total_snapshot_values = len(df['invalid snapshot value'].values.tolist())
invalid_snapshot_value_percentage = round((invalid_snapshot_value_count/total_snapshot_values) * 100, 2)

print(f'number of invalid snapshot values because of either a missing a snapshot value, duplicated snapshot value, or a missing family id is\n{invalid_snap
print(f'percentage of invalid snapshot values because of either a missing a snapshot value, duplicated snapshot value, or a missing family id is\n{invalid_s

invalid_snapshot_value_count = df['invalid snapshot value'].value_counts()[True]
total_snapshot_values = len(df['invalid snapshot value'].values.tolist())
invalid_snapshot_value_percentage = round((invalid_snapshot_value_count/total_snapshot_values) * 100, 2)

family_id_nan_count = df['Family Id'].isna().sum()
family_id_nan_percentage = round((invalid_snapshot_value_count/total_snapshot_values) * 100, 2)
print(f'number of missing family ids is {family_id_nan_count} out of {total_snapshot_values} total entries')
print(f'percentage of missing family ids is {family_id_nan_percentage}% out of {total_snapshot_values} total entries')
```

[29] ✓ 0.0s Python

... number of invalid snapshot values because of either a missing a snapshot value, duplicated snapshot value, or a missing family id is 1492 out of 654463 total entries
percentage of invalid snapshot values because of either a missing a snapshot value, duplicated snapshot value, or a missing family id is 0.23% out of 654463 total entries
number of missing family ids is 47 out of 654463 total entries
percentage of missing family ids is 0.23% out of 654463 total entries

6. Other Questions

1. How many families have transitioned from yellow to green per indicator?

1	number of familie that took at least one follow-up survey out of 6321 and transitioned from yellow to green per indicator are:
2	indicator capacityToPlanAndBudget had 1198 transitions
3	indicator diversifiedSourcesOfIncome had 607 transitions
4	indicator entrepreneurialSpirit had 697 transitions
5	indicator registeredToVoteAndVotesInElections had 444 transitions
6	indicator accessToCredit had 358 transitions
7	indicator accessToHealthServices had 319 transitions
8	indicator autonomyDecisions had 525 transitions
9	indicator awarenessOfHumanRights had 142 transitions
10	indicator clothingAndFootwear had 362 transitions
11	indicator phone had 428 transitions
12	indicator properKitchen had 317 transitions
13	indicator regularMeansOfTransportation had 445 transitions
14	indicator safety had 415 transitions
15	indicator accessInformation had 233 transitions
16	indicator documentation had 166 transitions
17	indicator influenceInPublicSector had 529 transitions
18	indicator vaccinations had 307 transitions
19	indicator abilityToSolveProblemsAndConflicts had 415 transitions
20	indicator selfEsteem had 582 transitions
21	indicator selfExpression had 73 transitions
22	indicator alimentation had 561 transitions
23	indicator awarenessOfNeeds had 956 transitions
24	indicator electricityAccess had 258 transitions
25	indicator garbageDisposal had 349 transitions
26	indicator householdViolence had 141 transitions
27	indicator middleEducation had 208 transitions
28	indicator readAndWrite had 285 transitions
29	indicator regularityOfMeals had 50 transitions
30	indicator unpollutedEnvironment had 622 transitions
31	indicator willingnessDesireToDevelopSkillsAndKnowledge had 90 transitions
32	indicator accessToShopsAndServices had 35 transitions
33	indicator culturalTraditionsAndHeritage had 80 transitions
34	indicator income had 611 transitions
35	indicator moralConscience had 326 transitions

1	number of familie that took at least one follow-up survey out of 6321 and transitioned from yellow to green per indicator are:
35	indicator moralConscience had 326 transitions
36	indicator refrigerator had 405 transitions
37	indicator respectForDiversity had 125 transitions
38	indicator safeHouse had 307 transitions
39	indicator sexualHealth had 354 transitions
40	indicator accessToEntertainment had 1000 transitions
41	indicator knowledgeAndSkillsToGenerateIncome had 708 transitions
42	indicator stableIncome had 68 transitions
43	indicator familySavings had 340 transitions
44	indicator dentalCare had 510 transitions
45	indicator groupActivities had 56 transitions
46	indicator eyesight had 437 transitions
47	indicator personalHygiene had 211 transitions
48	indicator safeBathroom had 348 transitions
49	indicator schoolSuppliesAndBooks had 206 transitions
50	indicator stableHousing had 63 transitions
51	indicator socialCapital had 896 transitions
52	indicator emotionalIntelligence had 493 transitions
53	indicator insurance had 457 transitions
54	indicator safetyFromFloods had 12 transitions
55	indicator securityOfProperty had 365 transitions
56	indicator comfortOfTheHome had 322 transitions
57	indicator separateBedrooms had 500 transitions
58	indicator childLabor had 102 transitions
59	indicator drinkingWaterAccess had 125 transitions
60	indicator floor had 50 transitions
61	indicator ageEducation had 32 transitions
62	indicator formalEmployment had 38 transitions
63	indicator unemployment had 37 transitions
64	indicator wall had 14 transitions
65	indicator addictions had 25 transitions
66	indicator childrenEducation had 8 transitions
67	indicator healthcareInfancy had 8 transitions

1	number of familie that took at least one follow-up survey out of 6321 and transitioned from yellow to green per indicator are:	
67	indicator healthcareInfancy had 8 transitions	
68	indicator debts had 246 transitions	
69	indicator road had 309 transitions	
70	indicator nearbyHealthPost had 258 transitions	
71	indicator covid_incomeSkills had 32 transitions	
72	indicator covid_internet had 33 transitions	
73	indicator covid_isolation had 31 transitions	
74	indicator covid_medicalAttention had 10 transitions	
75	indicator covid_nutrition had 3 transitions	
76	indicator covid_homeRoutine had 15 transitions	
77	indicator covid_supportNetwork had 9 transitions	
78	indicator covid_contingencyFunds had 22 transitions	
79	indicator covid_technology had 97 transitions	
80	indicator covid_choresDistribution had 10 transitions	
81	indicator covid_stableHouse had 4 transitions	
82	indicator covid_homeSupport had 11 transitions	
83	indicator covid_emotionalWellbeing had 8 transitions	
84	indicator covid_homeEducation had 53 transitions	
85	indicator domesticFertilizer had 4 transitions	
86	indicator knowledgeOfClimateChange had 9 transitions	
87	indicator plasticUsageReduction had 10 transitions	
88	indicator recycling had 11 transitions	
89	indicator reusesResidues had 5 transitions	
90	indicator sustainableTransportation had 8 transitions	
91	indicator energySavings had 8 transitions	
92	indicator waterConsumptionPatterns had 10 transitions	
93	indicator domesticWaterConsumption had 11 transitions	
94	indicator lowerEnergyItems had 10 transitions	
95	indicator promotesEnvironment had 4 transitions	
96	indicator accessToInternet had 73 transitions	
97	indicator growthMindset had 18 transitions	
98	indicator hopeInOpportunities had 25 transitions	
99	indicator familyPositiveBelonging had 20 transitions	
100	indicator safeOpportunities had 38 transitions	
101	indicator foodAccess had 31 transitions	
102	indicator emotionalSelfCare had 41 transitions	
103	indicator feelingOfWorthiness had 31 transitions	
104	indicator selfAcceptance had 38 transitions	
105	indicator debt had 21 transitions	
106	indicator helpingOthers had 30 transitions	
107	indicator rightsAndProtectionsKnowledge had 18 transitions	
108	indicator supportNetwork had 27 transitions	
109	indicator accessToRights had 25 transitions	
110	indicator dentalAndEyesightCare had 30 transitions	
111	indicator governmentBenefits had 38 transitions	
112	indicator safeWorkEnvironment had 11 transitions	
113	indicator skillsForSuccess had 32 transitions	
114	indicator socialPositiveBelonging had 34 transitions	
115	indicator accessToMentalHealthServices had 15 transitions	
116	indicator accessToMentors had 22 transitions	
117	indicator culturalFreedom had 33 transitions	
118	indicator wellbeingForThoseICareFor had 15 transitions	
119	indicator financialCompetence had 23 transitions	
120	indicator healthyLifestyle had 39 transitions	
121	indicator trust had 26 transitions	
122	indicator safeLifestyle had 23 transitions	
123	indicator bodyAutonomy had 19 transitions	
124	indicator freedomOfExpression had 30 transitions	
125	indicator spiritualWellbeing had 18 transitions	
126	indicator freedomAndHelpFromAbuse had 11 transitions	
127	indicator civicEngagement had 2 transitions	
128	indicator freedomFromSexualExploitation had 1 transitions	
129		

2. How many families have transitioned from red to green per indicator?

1	number of familie that took at least one follow-up survey out of 6321 and transitioned from red to green per indicator are:
2	indicator personalHygiene had 38 transition(s)
3	indicator vaccinations had 74 transition(s)
4	indicator stableIncome had 40 transition(s)
5	indicator abilityToSolveProblemsAndConflicts had 69 transition(s)
6	indicator accessToEntertainment had 362 transition(s)
7	indicator accessToHealthServices had 31 transition(s)
8	indicator autonomyDecisions had 76 transition(s)
9	indicator awarenessOfHumanRights had 35 transition(s)
10	indicator awarenessOfNeeds had 196 transition(s)
11	indicator capacityToPlanAndBudget had 643 transition(s)
12	indicator clothingAndFootwear had 54 transition(s)
13	indicator comfortOfTheHome had 67 transition(s)
14	indicator culturalTraditionsAndHeritage had 22 transition(s)
15	indicator dentalCare had 220 transition(s)
16	indicator documentation had 63 transition(s)
17	indicator electricityAccess had 31 transition(s)
18	indicator emotionalIntelligence had 57 transition(s)
19	indicator entrepreneurialSpirit had 103 transition(s)
20	indicator eyesight had 148 transition(s)
21	indicator familySavings had 314 transition(s)
22	indicator garbageDisposal had 265 transition(s)
23	indicator insurance had 290 transition(s)
24	indicator knowledgeAndSkillsToGenerateIncome had 197 transition(s)
25	indicator middleEducation had 63 transition(s)
26	indicator readAndWrite had 87 transition(s)
27	indicator regularMeansOfTransportation had 88 transition(s)
28	indicator regularityOfMeals had 17 transition(s)
29	indicator respectForDiversity had 17 transition(s)
30	indicator safeBathroom had 92 transition(s)
31	indicator safeHouse had 67 transition(s)
32	indicator safety had 67 transition(s)
33	indicator schoolSuppliesAndBooks had 48 transition(s)
34	indicator selfEsteem had 57 transition(s)
35	indicator selfExpression had 16 transition(s)

1	number of familie that took at least one follow-up survey out of 6321 and transitioned from red to green per indicator are:
35	indicator selfExpression had 16 transition(s)
36	indicator separateBedrooms had 165 transition(s)
37	indicator stableHousing had 21 transition(s)
38	indicator unpollutedEnvironment had 119 transition(s)
39	indicator willingnessDesireToDevelopSkillsAndKnowledge had 6 transition(s)
40	indicator influenceInPublicSector had 249 transition(s)
41	indicator moralConscience had 73 transition(s)
42	indicator income had 168 transition(s)
43	indicator groupActivities had 42 transition(s)
44	indicator socialCapital had 391 transition(s)
45	indicator accessInformation had 130 transition(s)
46	indicator accessToCredit had 318 transition(s)
47	indicator accessToShopsAndServices had 15 transition(s)
48	indicator alimentation had 102 transition(s)
49	indicator diversifiedSourcesOfIncome had 344 transition(s)
50	indicator householdViolence had 40 transition(s)
51	indicator phone had 112 transition(s)
52	indicator properKitchen had 136 transition(s)
53	indicator refrigerator had 117 transition(s)
54	indicator registeredToVoteAndVotesInElections had 105 transition(s)
55	indicator sexualHealth had 80 transition(s)
56	indicator safetyFromFloods had 7 transition(s)
57	indicator securityOfProperty had 94 transition(s)
58	indicator road had 73 transition(s)
59	indicator drinkingWaterAccess had 27 transition(s)
60	indicator healthcareInfancy had 8 transition(s)
61	indicator childrenEducation had 4 transition(s)
62	indicator formalEmployment had 19 transition(s)
63	indicator unemployment had 11 transition(s)
64	indicator ageEducation had 3 transition(s)
65	indicator addictions had 9 transition(s)
66	indicator physicalActivity had 2 transition(s)
67	indicator childLabor had 38 transition(s)
68	indicator nearbyHealthDnct had 56 transition(s)

3. Which indicator had the most favorable changes from yellow to green?

```
# which indicator had the most favorable changes from yellow to green
max_count = None
max_indicator = None

total_num_of_families = len(family_ids)
for indicator, count in indicator_transition_yellow_to_green.items():
    if max_count is None:
        max_count = count
        max_indicator = indicator
    elif max_count < count:
        max_count = count
        max_indicator = indicator

print(f'the indicator that had the most favorable changes from yellow to green is {max_indicator} with {max_count} transitions out of {total_num_of_families}
```

[60] ✓ 0.0s Python

... the indicator that had the most favorable changes from yellow to green is capacityToPlanAndBudget with 1198 transitions out of 6321 total number of families who took at least one follow-up survey

4. Which indicator had the most favorable changes from red to green?

```
# which indicator had the most favorable changes from red to green
max_count = None
max_indicator = None

total_num_of_families = len(family_ids)
for indicator, count in indicator_transition_red_to_green.items():
    if max_count is None:
        max_count = count
        max_indicator = indicator
    elif max_count < count:
        max_count = count
        max_indicator = indicator

print(f'the indicator that had the most favorable changes from red to green is {max_indicator} with {max_count} transitions out of {total_num_of_families} t
```

[61] ✓ 0.0s Python

... the indicator that had the most favorable changes from red to green is capacityToPlanAndBudget with 643 transitions out of 6321 total number of families who took at least one follow-up survey

5. What is the average time between the baseline survey and the follow-up survey?

```
# what is the average time between the baseline survey and the follow-up survey
family_count_with_follow_ups = len(last['Family Id'].values.tolist())
average = round((total_time_difference_in_weeks/4/family_count_with_follow_ups, 2)
print(f'the average time between the baseline survey and the follow-up survey is {average} months')
```

[35] ✓ 0.0s Python

... the average time between the baseline survey and the follow-up survey is 11.45 months