Survey Data Cleaning

viwa

2024-08-22

1. Cleaning Post-Season Survey Data

Post-Season survey data was collected via email, social media, and other online methods. As a result, despite reCAPTCHA inclusion, there are high numbers of bot responses within the data set. Initial survey analysis requires cleaning and removing these responses from the dataset.

Cleaning using IP Address methods. A separate .csv file on CrowdSignal collected participant data for the post-season survey. This will be cross-joined with the post-season survey data set on primary key "Respondent.ID" to determine the IP Addresses for all respondents. Then, evidence-based IP analysis can determine bot presence.

First 6 rows of joined table data containing both survey response and participant data.

Respondent ID	IP Address
304883026	120.230.121.13
304892255	104.233.228.182
304790849	80.188.22.93
304713215	176.121.238.148
304713306	176.121.248.159
304723120	176.121.238.214

Grouped by IP.Address in order to count multiple responses for same IP addresses. Because count higher than 1 indicate more than one response, we want to filter.

First 6 rows after removing responses from duplicated IP Addresses

As we can see, there are still bots present that will need to be removed through additional methods.

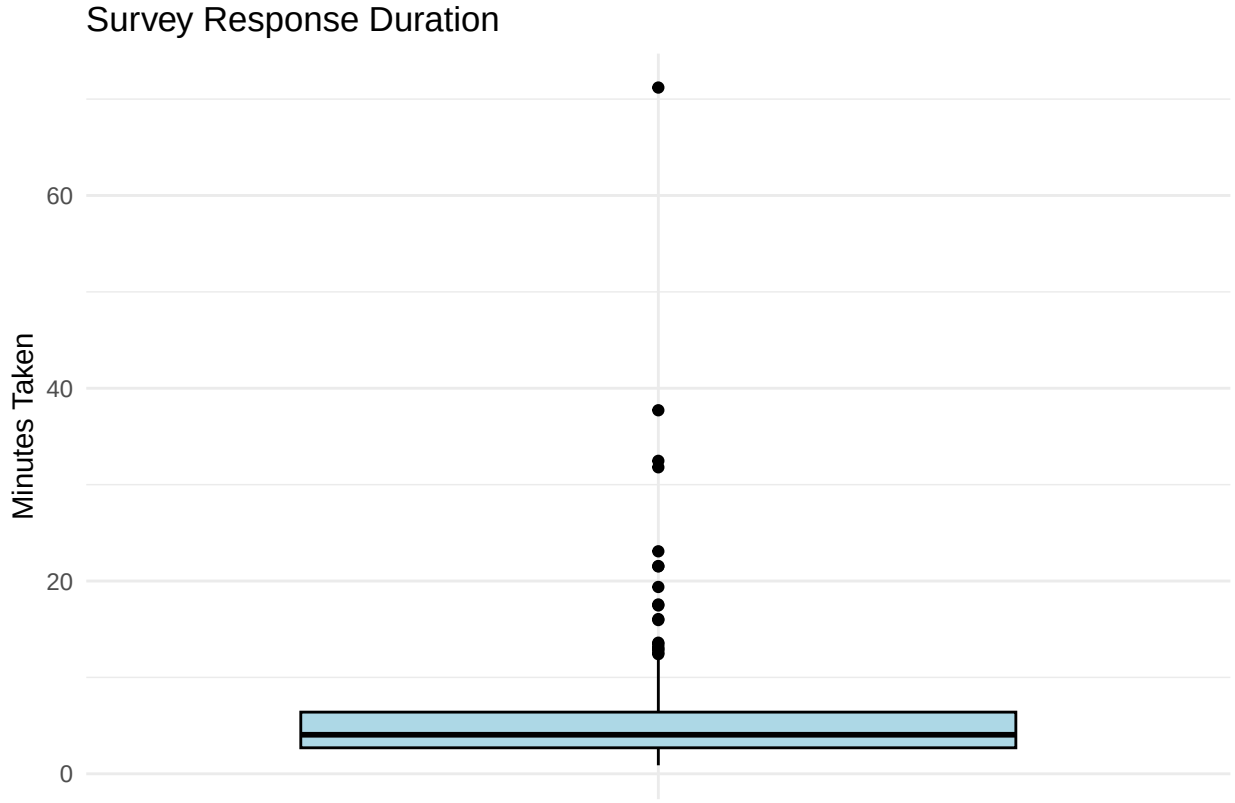
Response ID	IP Address	Anything else to share?
304883026	120.230.121.13	Top of the Park or A2SF, as an organization, is committed to promoting the sustainable development of parks and public spaces and improving the quality of urban life. They promote the design, construction, and maintenance of parks and public spaces through advocacy, education, and collaboration to provide better places for citizens to enjoy leisure, sports, and socialize. In addition, A2SF also focuses on the role of parks and public spaces in environmental protection, climate change and community development, and promotes relevant policies and
304892255	104.233.228.182	I really enjoy these activities and things like that

304790849	80.188.22.93	As a first time visitor to A2SF, I was really impressed with the level of musicians and other artists, the food and beer tent. It was clean and had a good vibe.
304713215	176.121.238.148	no
304713306	176.121.248.159	no
304723120	176.121.238.214	Enjoy the music atmosphere

We also want to filter on all other selections to double check we aren't losing responses. One address was identified with two responses, but one missed the email address while the other contained one. We write the likely_bots to a new file, "duplicate_ip_address.csv" which is provided in the code file.

Filtering by unusually short survey response duration. *Using the boxplot method, determine unusually short survey response durations.*

Following IP verification, we can look further into specific responses that may have evaded this method. Another strong indicator of bots are unusually short response durations. Because Time.Taken is a parameter within the post-season-respondents (participants) survey, we can use these numbers to determine unusually short responses.



Takeaways: Because only some responses were unusually long, there is no evidence for bots present based on extremely short response times only. However, after writing post_season_ip to a new file, "post_season_cleaned_ip_addresses.csv" there are clearly still bots within the data.

Top 6 fastest response durations for post-season respondents.

Respondent ID	Duration (min:sec)
304720498	0:53
304713306	0:54
304713284	1:00
304713215	1:11
304724764	1:14
304885124	1:21

Combining manual- and auto-collected parameters to validate participant responses. *Remove non-US auto-collected countries and manual zip code entries that are not 5 characters.*

Rationale: Auto-collected countries outside of the United States are unlikely given the scope and target audience of the survey. Additional manual review of row responses for non-US locations indicated bot behavior such as nonsensical free response answers.

All non-US Post-Season Responses (n = 10).

Non-US countries should not have filled out the zip code question. The "Anything to share" column is used as an additional verification of bot-indicating behavior. As a result of this, we can filter these rows out of the dataset further.

Respondent ID	Country	Zip Code	Anything else to share?
304883026	China	10004	Top of the Park or A2SF, as an organization, is committed to promoting the sustainable development of parks and public spaces and improving the quality of urban life. They promote the design, construction, and maintenance of parks and public spaces through advocacy, education, and collaboration to provide better places for citizens to enjoy leisure, sports, and socialize. In addition, A2SF also focuses on the role of parks and public spaces in environmental protection, climate change and community development, and promotes relevant policies and
304892255	China	48104	I really enjoy these activities and things like that
304790849	Czech Republic	48381	As a first time visitor to A2SF, I was really impressed with the level of musicians and other artists, the food and beer tent. It was clean and had a good vibe.
304713215	Ukraine	48105	no
304713306	Ukraine	48107	no
304723120	Ukraine	7104	Enjoy the music atmosphere
304723498	Ukraine	6915	Enjoy the atmosphere that music brings to me
304741418	Ukraine	72701	Rooftop Interior Room: A 700-plus square foot interior space that can accommodate up to 70 guests with year-round views of the city, perfect for family gatherings and small weddings

304741907	Ukraine	48104	The warmth and happiness of spending happy time with your family
304745032	Ukraine	67758	Food booth: Offering a variety of food and drinks to meet the needs of different tastes

Additionally, zip codes should be 5 characters in length. When we see typos like "4103" instead of "48103", it is possible this is a typo, but we should not assume this to be a zip code in 48103. Further filtering requires the removal of these responses.

Respondent ID	Country	State	Zip Code
304885041	United States of America	Arizona	renegofdorsey@gmail.com
304885502	United States of America	California	713499
304911308	United States of America	California	389-12-9046
304681059	United States of America	Michigan	48103-1736
304885727	United States of America	New York	Deb Reese
304911852	United States of America	New York	5051
304725609	United States of America	Pennsylvania	6042
304911504	United States of America	Texas	Leopold Burns

NOTE: Response ID 304681059 seems to be a valid combination of Country, State, and Zip Code, and upon reviewing the other responses, indicates a human responded. This entry is edited and will remain in the dataset, but with only the first 5 digits (48103) instead of including -1736 in the end.

Overall, zip code entries that are not 5 characters are unlikely, and few countries outside the US use postal codes.

Determine bots using zip code and state collected data.

Rationale: Because some United States-based responses may also be bots, we have to be creative to identify additional ones. To do this, I use a function to identify states and zip codes and determine whether the zip codes, which respondents manually inputted, match with the states/countries, which was automatically collected from respondents' device locations.

First six respondent ID's and their state-zip code matching.

Following identification of "Match" or "Not Match", we filter the full dataset so it only contains "Match" results. Then, we will write to csv a fully cleaned file that can be used in further post-season analysis and data visualization methods.

Respondent ID	State	Zip Code	Match?
304882026	Alabama	99648	No Match
304884161	Alabama	59601	No Match
304884684	Alabama	99550	No Match
304885124	Alabama	35244	Match
304887014	Alabama	65762	No Match
304883642	Arizona	48109	No Match

Final cleaned result. There are still some suspicious responses because the short answer does not align well with the question asked, but these are few in number.

Notably suspect: **304885124, 304722909, 304882174, 304751759**. However, these managed to pass ALL of the following:

- Non-duplicate IP Address sources

- Normal response duration
- United States
- Zip code length is 5 characters
- State (auto-collected) and Zip code (manual-collected) match

As a result, we will leave these responses in the dataset, but should exercise caution when performing analysis.

ID	State	Zip Code	Anything else?	Email	Match?
304885124	Alabama	35244	Recruit volunteers to guide the audience	janzotell82@gmail.com	Match
304885234	Arizona	85138	Invite more established performers	ratherp04@gmail.com	Match
304882174	California	90017	Hope to provide more services	myerstims1999@gmail.com	Match
304751759	Louisiana	70112	Let me share the content of the event Activity Content: Free Concert: Every night, different styles of music are performed, from local bands to internationally renowned artists. Movies by Moonlight: Outdoor movie screenings for families and friends. Street performance: A variety of open street performances and art displays. Food booth: Offering a variety of food and drinks to meet the needs of different tastes. Family Activities: A variety of interactive activities and games for families and children.	laderoutenicholai@gmail.com	Match
304722909	Massachusetts	13130	A good musical experience is one of passion and creativity	marinogeronimo839@gmail.com	Match
304677741	Maryland	21209	Thank you! We were at Umich for a conference and decided to stop by.	heshyr@gmail.com	Match

The final output of this file is named "post_season-processed.csv".