Stanley Park Survey - Response to COVID-19

Data Cleaning Report

November 2020

1. Data Cleaning

Data cleaning is the process of detecting and fixing (or removing) inaccurate, incomplete, duplicate or corrupt records from a dataset. Data cleaning is conducted to ensure the accuracy and completeness of the data.

2. Data Cleaning Process

A multi-step approach was taken to clean the survey data (Table 1). Decisions to remove cases from the survey data were based on judgements about these criteria. Records were not removed for simply having the same IP address, as there are legitimate reasons for this (such as having the same residential or work address).

Table 1: Data cleaning criteria and actions

Criteria	Action	Method	Results
Duplicate IP addresses	Flag & check for suspicious behaviours. Remove if identical responses are found.	Excel conditional formatting. Analyses to determine effect of duplicates on survey results (see below).	2,074 records
Identical responses	Remove if same IP address. Remove if open responses are identical as well.	SPSS Identify Duplicate Cases	None found
Speeders Respondents who complete the survey in an unrealistically short time.	Flag if under 40% of the median survey duration. Remove if under 30% of median duration. 30-40% was considered reasonable due to survey branching (question skipping).	Calculate time taken. Compare with median survey duration length	Median survey duration is 10:41 (mm:ss). Survey completion <30% of median 3:12 (mm:ss) were removed. 187 cases removed.
Submit time (For IP duplicates) Mainly to see if respondents submitted their surveys more than once (e.g. due to poor internet connection).	Flag if identical submit time (within 3 minutes) from same IP) Flag if multiple similar submit times from same IP address	Excel filtering (IP address and submit time)	91 found from duplicate IP addresses. A decision was made to keep these, since no identical responses were found, nor was there other suspicious activity.

Criteria	Action	Method	Results
Straight Liners Identify respondents who always provide the same response (e.g. always select the first option). Interpret straight liners with caution for agree/disagree statements – responses may be legitimate.	Flag	Apply filters in Excel Check if respondents always answer the first option.	No suspicious activity found
Nonsensical open-ended responses	Flag	Manual scan	None found
Inconsistencies Check for survey responses that contradict each other. Compare responses to 'Don't or Didn't visit' SP questions - with caution since people can be inconsistent (e.g. ambiguity/ misunderstand a question in a different context).	Flag	Excel apply multiple filters	62 found; 8 found from duplicate IP addresses. Ignored these as there was no other suspicious activity.
Total cases removed:		**187**	

3. Duplicate IP analysis

There were responses to the survey from respondents with the same IP address. This can arise when different members of the same household or place of work (who would have the same IP address) complete the survey. With web-based surveys, there is also a potential risk of respondents completing this survey more than once. As part of the data cleaning process, we determined the effects of responses from duplicate IP addresses on the survey results. The following analyses were conducted.

3.1 Comparison of Duplicate and Non-duplicate groups

For the first analysis, responses with and without duplicate IP addresses were treated as two separate groups, and cross-tabulations were computed in SPSS for each survey question, for both groups. The percent of respondents selecting each response for the 'Duplicate' and 'Non-duplicate' groups were compared, and the difference between the groups was calculated for each survey question.

The Minimum, Maximum, Mean, and Standard Deviation were then computed for each group, and the difference between the two groups was also calculated (Table 2). The largest percentage of difference for any question was 4.3%. The average was 0.96%. For most survey questions (95%), the percentage difference was between 0% and 2%. Table 3 shows the number and proportion of survey questions for each percentage difference. Table 4 shows the survey questions that had the highest (3% to 4%) difference.

Table 2: Difference between 'Duplicate' and 'Non-duplicate' groups

Survey Question (% Yes)	Duplicate IP	Non-duplicate	Difference
Min	0.10%	0.10%	0.00%
Max	82.60%	82.30%	4.30%
Mean	19.89%	19.84%	0.96%
Standard Deviation	0.1943286	0.190438406	0.008963

Table 3: Number and proportion of survey questions for each percentage difference between the two groups

Percentage difference	Number of survey questions	%
0%	81	41%
1%	63	32%
2%	42	21%
3%	7	4%
4%	3	2%

Table 4: Survey responses with 3% or 4% difference between the two groups

Survey question	Response with 3% or 4% difference
If you experienced Stanley Park and felt it was better when it was temporarily closed to vehicles, let us know why.	'I found it more quiet and peaceful'
How did you travel to get to Stanley Park when it was re-opened with one lane for cars and one lane for bikes?	Walk/Run
Why didn't you use the cycle lane?	I preferred the seawall cycling path over the road
You've said you didn't visit Stanley Park since it was opened to vehicles on June 22nd, why is that?	I was concerned about being exposed to COVID-19 AND I typically like to drive to the Park and avoided it due to one lane being dedicated to cyclists
How often did you visit and use the Park when it was re-opened with one lane for cars and one lane for bikes compared to when it was car-free?	The same
If you have visited Stanley Park since it was re-opened to vehicles on June 22nd, how was your Park experience compared to when it was closed to vehicles?	Worse than when it was closed to vehicles
Zone_Rollup	n/a

3.2 Comparison of the whole sample with and without duplicates

In the second analysis, two datasets were created; one including the responses from duplicate IP addresses ('With Duplicates') and one excluding them ('Without Duplicates'). To compare the responses to each survey question between the two datasets, percentages were calculated for each survey item, and the difference between them was determined (Table 5). The largest difference between the two datasets for any survey question was 0.8%, and the average (mean) was -0.002% difference.

Table 5: Comparison of the whole sample with and without duplicates

	With Duplicates	Without Duplicates	Difference
MIN	0.0%	0.0%	-0.80%
MAX	99.9%	99.9%	0.80%
Mean	41.3%	41.3%	-0.002%
SD	0.356771411	0.356534251	0.002495

From these analyses, we can conclude that the inclusion of survey responses from duplicate IP addresses has a small effect on survey results.