COMP3839 DATA QUALITY IMPROVEMENT

Anjali Vekaria

Contents

Overview of Project

For my final project, I've chosen the 2018 dataset focusing on business licenses in Vancouver. The goal is to explore the available data and extract meaningful insights. This dataset is intriguing, boasting over 66,000 records and 25 columns detailing licenses and business renewals in the city. In Vancouver, having a valid business license is a must, and these licenses are obtained from the City's License Office. They remain active for the entire current calendar year unless stated otherwise. One crucial column, 'LicenseRSN,' uniquely identifies businesses with numerical values, and tools like DQAnalyzer aid in a more comprehensive understanding. Another vital column, 'BusinessName,' uses strings to accommodate the varied names businesses might have, ensuring a complete representation. The table below shows all the columns along with a short description. The data is used from the https://opendata.vancouver.ca/explore/dataset/business-licences/

Column Name	Description
FolderYear	The year in which the business license data is recorded.
LicenceRSN	A unique identification number for each license record.
LicenceNumber	The specific license number assigned to a business.
LicenceRevisionNumber	Number indicating the revision of the license, if applicable.
BusinessName	The legal name of the business entity.
BusinessTradeName	Any trade name or DBA (Doing Business As) name used by the business.
Status	The status of the business license (e.g., active, expired).
IssuedDate	The date when the license was issued.
ExpiredDate	The date when the license is set to expire.
BusinessType	The broad category or industry in which the business operates.
BusinessSubType	A more specific category within the broader business type.
Unit	The unit number if the business is in a multi-unit building.
UnitType	The type of unit (e.g., suite, apartment).
House	House number of the business address.
Street	The street name where the business is located.
City	The city where the business is situated (Vancouver in this case).
Province	The province of the business location.
Country	The country where the business operates.
PostalCode	The postal code of the business address.
LocalArea	The specific local area or neighborhood within Vancouver.
NumberOfEmployees	The count of employees working for the business.
FeePaid	The amount of fee paid for the license.
ExtractDate	The date when the data was extracted from the source.
Geom	Geospatial data representing the geometry of the business location.
Geo_point_2d	Latitude and longitude coordinates of the business location.

As a data analyst my role would be find out the issues from the records based on the columns. Here I will be responsible to showcase what I find needs to be drawn attention to based on the results generated from DQAnalyzer. As a technical data analyst, would be handling the technical aspects of data processing to ensure the accuracy and reliability of the dataset. The fictional SME I am portraying is responsible in analyzing the issues forwarded by data analyst and help resolve the issue.

My approach for the project would be based on following parameters:

Initial Assessment: In the initial assessment, the technical data analyst evaluates the 2018 business licenses dataset from the City of Vancouver for potential insights.

SME Review of Initial Assessment: The Subject Matter Expert (SME) critically examines the initial assessment findings, validating the methodology used and ensuring the analysis aligns with the dataset's context and nuances.

Further Research for the SME: Additional in-depth investigation conducted by the technical analyst based on the SME's feedback and specific queries, aiming to explore deeper insights.

SME Review of Further Research: The SME reviews the results of the additional research, assessing the new findings and determining their relevance and accuracy in enhancing the understanding of the dataset.

SME Suggests Some DQ Rules: The SME proposes specific Data Quality (DQ) rules, outlining guidelines and standards to validate and enhance the quality of the dataset, ensuring accurate, consistent, and reliable information for analysis and decision-making.

Initial Assessment

LicenceRSN : Duplicates

The LicenceRSN column is a unique identifier that helps to identify businesses which are presumed to be unique but has 0.04% duplicate data as shown in the below image. There are 25 records in the dataset which has values duplicated making it difficult to analyze.

Type	Count	%
Null	0	0.00%
Non-null	66,181	100.00...
Duplicate	25	0.04%
Distinct	66,156	99.96%
Non-uniq...	25	0.04%
Unique	66,131	99.92%

The image below shows that in the count of 2 the LicenseRSN has been repeated. The blue colored from top shows the same LicenseRSN but have different PostalCodes. Also, at a glance we could see the postalcodes are different for same LicenseRSN and has either pending/cancelled or issued status.

LicenceRSN	BusinessName	Status	City	PostalCode
2995807	Sterum Properties Ltd	Pending	Vancouver	V6B 1S2
2995807	Sterum Properties Ltd	Pending	Vancouver	V6G 2C7
2999828	Sang Eun Lee (Sang Lee)	Cancelled	Vancouver	V6E 4R5
2999828	Sang Eun Lee (Sang Lee)	Cancelled	Vancouver	V6T 1R3
2999829	Sang Eun Lee (Sang Lee)	Gone Out of Business	Vancouver	V6E 4R5
2999829	Sang Eun Lee (Sang Lee)	Gone Out of Business	Vancouver	V6T 1R3
3003582	1003259 BC Ltd	Pending	Vancouver	
3003582	1003259 BC Ltd	Pending	Vancouver	
3003583	1003259 BC Ltd	Cancelled	Vancouver	
3003583	1003259 BC Ltd	Cancelled	Vancouver	
3012422	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012422	Vancouver City Savings Credit Union	Issued	Vancouver	V6M 2A4
3012423	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012423	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012424	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012424	Vancouver City Savings Credit Union	Issued	Vancouver	V6K 1N9
3012425	Vancouver City Savings Credit Union	Issued	Vancouver	V6R 2B1
3012425	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012426	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012426	Vancouver City Savings Credit Union	Issued	Vancouver	V5L 3Y3
3012427	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012427	Vancouver City Savings Credit Union	Issued	Vancouver	V5K 1Z3
3012428	Vancouver City Savings Credit Union	Issued	Vancouver	V5R 5K6
3012428	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012429	Vancouver City Savings Credit Union	Issued	Vancouver	V5V 3P8
3012429	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012430	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012430	Vancouver City Savings Credit Union	Issued	Vancouver	V5Z 1K9
3012431	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012431	Vancouver City Savings Credit Union	Issued	Vancouver	V6S 2G4
3012432	Vancouver City Savings Credit Union	Issued	Vancouver	V5W 3A4
3012432	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012433	Vancouver City Savings Credit Union	Issued	Vancouver	V6C 1J8
3012433	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012434	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012434	Vancouver City Savings Credit Union	Issued	Vancouver	V5P3W1
3012557	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012557	Vancouver City Savings Credit Union	Issued	Vancouver	V6R 4N2
3012672	Vancouver City Savings Credit Union	Issued	Vancouver	V6A 4G2
3012672	Vancouver City Savings Credit Union	Issued	Vancouver	V6G 1C7
3178764	(Keiko Yokoyama)	Issued	Vancouver	V6Z 2Y7
3178764	Iemitsu Yokoyama & Keiko Yokoyama	Issued	Vancouver	V6Z 2Y7
3179278	Scott Sheng Heng Chan (Scott Chan)	Cancelled	Vancouver	V5L 1P4
3179278	Scott Sheng Heng Chan (Scott Chan)	Cancelled	Vancouver	V5Y 1G4
3196291	(Daisy Leo)	Pending	Vancouver	V6B 0G4
3196291	Lawrence Leo & Daisy Leo	Pending	Vancouver	V6B 0G4
3225479	0708306 BC Ltd	Pending	Vancouver	V6P 4R2
3225479	0708306 BC Ltd	Pending	Vancouver	V6J 1T5
3310148	(Suyeon Yu)	Pending	Vancouver	V6E 1H8
3310148	(Suyeon Yu)	Pending	Burnaby	V5H 1S9

BusinessName : Null

The BusinessName column indicates the name of each business and almost 5.6% which is 3753 businesses do not have a name and are null. The businesses are registered with the licence but do not have a name.

Type	Count	%
Null	3,753	5.67%
Non-null	62,428	94.33%
Duplicate	8,725	13.18%
Distinct	53,703	81.15%
Non-uniq...	5,122	7.74%
Unique	48,581	73.41%

Here we have taken BusinessName=Null and considered all the BusinessTradeName for image on the left and counted the status for each of the status. The number 3753 matches to our basic check performed in the DQ Analyzer.

For the image on the right, we have considered BusinessName=Null and BusinessTradeName=NULL and found the differnce of 1 in the inactive status which resulted in total 3752 for all the status. As we can see the status=issues is greater than 100, so we can further drill down by counting business type to understand the pattern.

BusinessName	(blank)
BusinessTradeName	(All)

Status	Count of Status
Cancelled	273
Gone Out of Business	51
Inactive	21
Issued	3250
Pending	158
Grand Total	**3753**

BusinessName	(blank)
BusinessTradeName	(blank)

Status	Count of Status
Cancelled	273
Gone Out of Business	51
Inactive	20
Issued	3250
Pending	158
Grand Total	**3752**

BusinessName	(blank)
BusinessTradeName	(blank)

Status	Count of BusinessType
Issued	**3250**
Electrical Contractor	1
One-Family Dwelling	1
Short-Term Rental	3248
Grand Total	**3250**

PostalCode: Invalid Format

In this dataset 36,675 are non null records and 29,506 are null records. A postal code is a part of an address that identifies a specific location to deliver mail.

Type	Count	%
Null	29,506	44.58%
Non-null	36,675	55.42%
Duplicate	30,891	46.68%
Distinct	5,784	8.74%
Non-uni...	3,375	5.10%
Unique	2,409	3.64%

Postal code is a series of letters and/ or digits that is attached to the address, but here in this dataset some postal codes are in invalid format. Postal code should not be in invalid format because businesses receive a lot of mail every day and wrong postal code could deliver it at wrong location.
All records have a postal code in correct format of LDL DLD, 99999, 99999-9999. Some data in the postal code column are only Letter(L) or Digit(D) or incorrect format. The highlighted blue here shows the correct format for postal code.

Mask Analysis

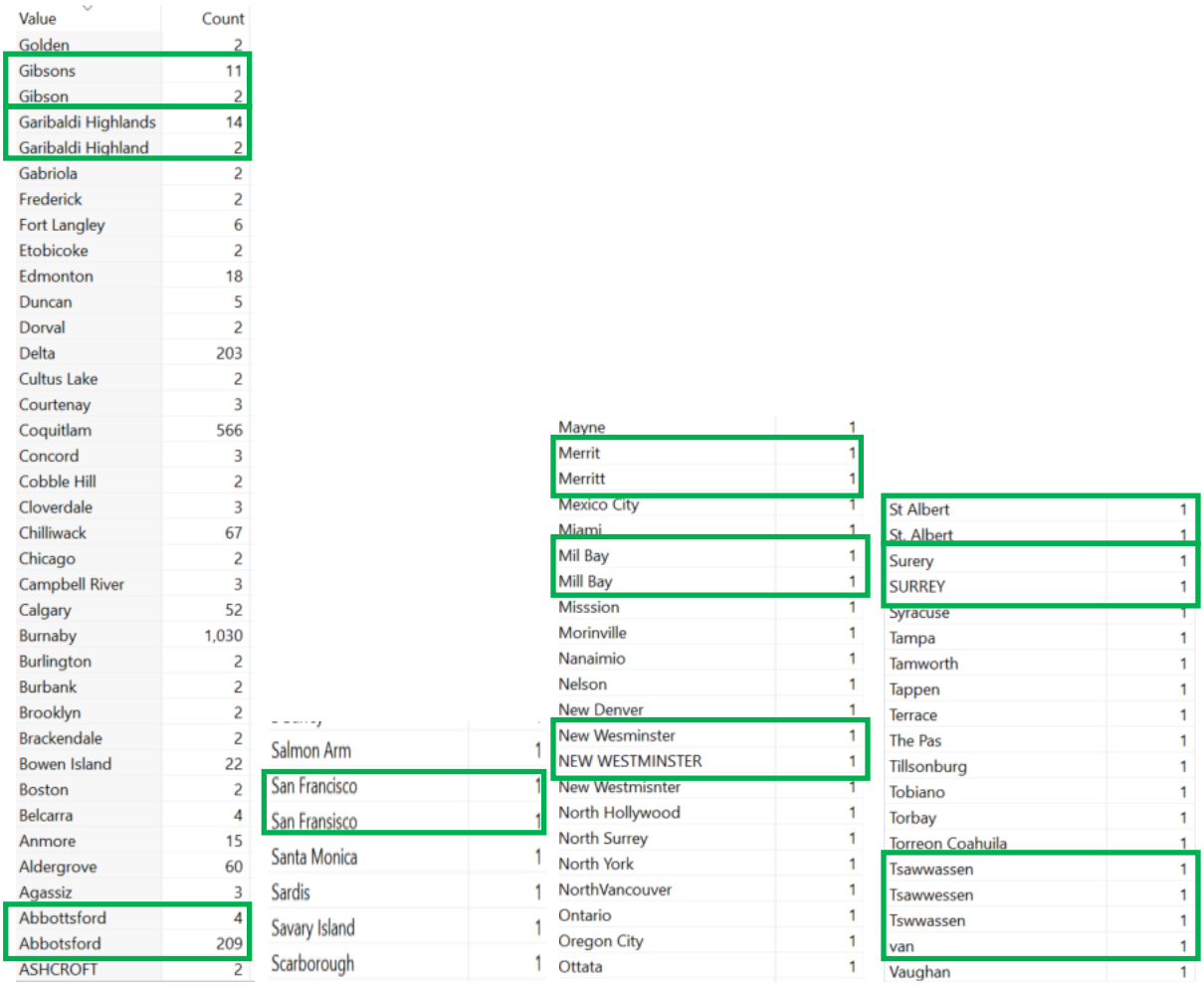
Mask: characters: [:letter:] -> L,[:digit:] -> D

Value	Count	%
NULL	29,506	44.58%
LDL DLD	36,116	54.57%
LDLDLD	384	0.58%
LDL DLD	77	0.12%
LDL DLL	17	0.03%
LLL LLLLLL	10	0.02%
LDD DLD	8	0.01%
LDL DDD	8	0.01%
DDDDD	6	0.01%
LDL LLD	6	0.01%
LDL DDL	4	0.01%
LL	4	0.01%
L/L	3	0.00%
LD DLD	3	0.00%
LDL	3	0.00%
L	2	0.00%
LD LDLD	2	0.00%
LDL)LD	2	0.00%
LDL DLD`	2	0.00%
LDLL DLD	2	0.00%
LLL DLD	2	0.00%
DDDDDD	1	0.00%
DDDDDDDDDD	1	0.00%
DDL DLD	1	0.00%
DLLDLD	1	0.00%
LD DLL	1	0.00%
LD: DLD	1	0.00%
LD& DLD	1	0.00%
LDL DD	1	0.00%
LDL DLD\L\L	1	0.00%
LDL DLDDDDDD	1	0.00%
LDLDDD	1	0.00%
LLD DLD	1	0.00%
LLDDDDDDD	1	0.00%
LLDL DLD	1	0.00%

City: Mis-spellings

The 2018 data set has some misspellings as well as nulls for City column. Misspellings in the 'City' column might cause data inconsistencies. These errors, such as variations in city names or typographical mistakes, can impact data accuracy and need attention during analysis to ensure reliable results. The table below shows there are 83 nulls in the city column.

The following frequency analysis table shows some mis-spellings that are found in the city column.



Value	Count
Golden	2
Gibsons	11
Gibson	2
Garibaldi Highlands	14
Garibaldi Highland	2
Gabriola	2
Frederick	2
Fort Langley	6
Etobicoke	2
Edmonton	18
Duncan	5
Dorval	2
Delta	203
Cultus Lake	2
Courtenay	3
Coquitlam	566
Concord	3
Cobble Hill	2
Cloverdale	3
Chilliwack	67
Chicago	2
Campbell River	3
Calgary	52
Burnaby	1,030
Burlington	2
Burbank	2
Brooklyn	2
Brackendale	2
Bowen Island	22
Boston	2
Belcarra	4
Anmore	15
Aldergrove	60
Agassiz	3
Abbottsford	4
Abbotsford	209
ASHCROFT	2
Salmon Arm	1
San Francisco	1
San Fransisco	1
Santa Monica	1
Sardis	1
Savary Island	1
Scarborough	1
Mayne	1
Merrit	1
Merritt	1
Mexico City	1
Miami	1
Mil Bay	1
Mill Bay	1
Misssion	1
Morinville	1
Nanaimio	1
Nelson	1
New Denver	1
New Wesminster	1
NEW WESTMINSTER	1
New Westmisnter	1
North Hollywood	1
North Surrey	1
North York	1
NorthVancouver	1
Ontario	1
Oregon City	1
Ottata	1
St Albert	1
St. Albert	1
Surery	1
SURREY	1
Syracuse	1
Tampa	1
Tamworth	1
Tappen	1
Terrace	1
The Pas	1
Tillsonburg	1
Tobiano	1
Torbay	1
Torreon Coahuila	1
Tsawwassen	1
Tsawwessen	1
Tswwassen	1
van	1
Vaughan	1

Province: 2-Capitalized Letter Standard Not Followed

The 2018 dataset has 90 null values in Province column. A notable issue arises with the 'Province' column where the standard of using two capitalized letters for province abbreviations is not consistently followed. This inconsistency, such as mixing uppercase and lowercase letters, can lead to confusion and hinder data uniformity.

Type	Count	%
Null	90	0.14%
Non-null	66,091	99.86%
Duplicate	66,052	99.81%
Distinct	39	0.06%
Non-uni...	24	0.04%
Unique	15	0.02%

The frequency analysis below shows the inconsistency in province column.

Value	Count
NULL	90
	1
YT	1
WA	12
Vancouver	4
VA	1
TX	2
SK	2
SC	1
QC	16
PQ	5
PA	1
On	3
OR	1
MA	5
Il	1
IN	1
IL	1
FL	3
DE	2
CT	2
CO	2
CA	30
British Columbia	22
BC	65,746
Ab	2
AZ	1
AL	1
AB	73
78	1

SME Review of Initial Assessment

LicenceRSN: Duplicates (High)

The analysis for LicenseRSN shows some of the duplicates with a high priority as this is one of the most important pieces of information that would eventually help us to keep correct information for each business. Probably can check the frequency of the duplicates and try to understand the pattern that led to duplicates.

BusinessName : Null (Medium)

The null values in the BusinessName column needs to be addressed on a medium priority as the businesses needs to be assigned names that would be useful for people to find it and even easy to address if the issue occurs. A combination of LicenceRSN and BusinessName would be a ideal way to spot information about that particular business.

PostalCode: Invalid Format (Medium)

PostalCode is vital information for the business and having address with correct postal code is needed. The 'PostalCode' column has been flagged with a medium-priority concern due to invalid formatting. Rectifying these invalid formats with medium priority is crucial to ensure consistency and accuracy in postal code data.

City: Mis-spellings (Medium)

From the SME's standpoint, the 'City' column presents a medium-priority issue due to misspellings. While not critical, these inaccuracies could introduce confusion and compromise data reliability. Given medium priority, it's essential to rectify these mis-spellings to enhance the dataset's accuracy.

Province: 2-Capitalized Letter Standard Not Followed (Medium)

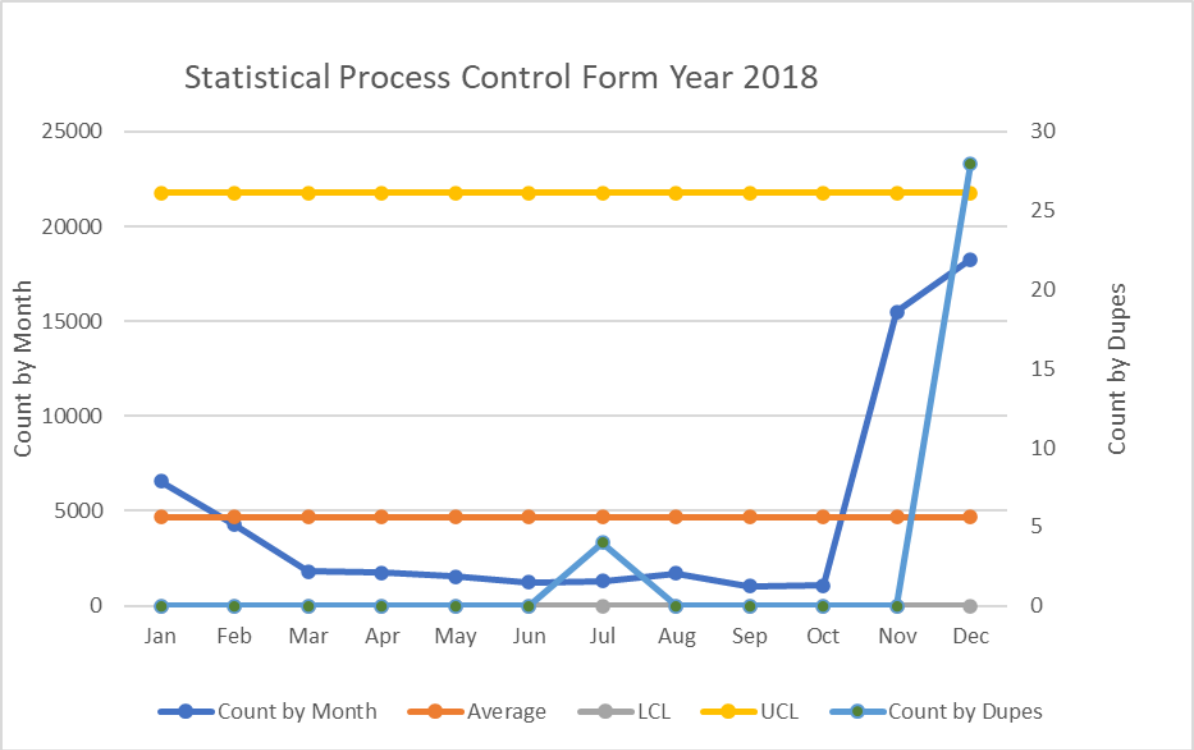
Regarding the 'Province' column, the SME has identified a medium-priority concern due to the inconsistent use of two capitalized letters. While not urgent, this issue could lead to confusion and impact data uniformity. Addressing this matter with medium priority is essential to uphold standard formatting, ensuring accurate representation of provinces for reliable analysis and decision-making.

Further Research for the SME

The SME recommended using 'LicenceRSN' as the unique identifier but identified duplicates in the data. By leveraging issue dates, we created a pivot table to filter the number of licenses issued monthly in 2018. We even calculated duplicates by months. Null 'IssuedDate' records were not considered. We calculated the average, lower limit, and upper limit for this data, and visualized it through a statistical process control chart. Count by dupes shows duplicates entries based on month. Notably, all counts fell within the upper and lower control levels, indicating that the process was statistically stable and under control.

Status	(All)				
Row Labels	**Count by Month**	**Count by Dupes**	**Average**	**LCL**	**UCL**
Jan	6550	0	4662	0	21785
Feb	4273	0	4662	0	21785
Mar	1794	0	4662	0	21785
Apr	1729	0	4662	0	21785
May	1524	0	4662	0	21785
Jun	1231	0	4662	0	21785
Jul	1297	4	4662	0	21785
Aug	1703	0	4662	0	21785
Sep	1014	0	4662	0	21785
Oct	1068	0	4662	0	21785
Nov	15491	0	4662	0	21785
Dec	18271	28	4662	0	21785

Form Type	**Average**	**Std Dev.**
2018	4662	5708



SME Review of Further Research

The process is confirmed to be under statistical control based on the 2018 data. This conclusion is drawn from the fact that all counts consistently fell within the predetermined upper and lower control limits, indicating a stable and predictable pattern in the dataset.

SME Suggests Some DQ Rules

LicenceRSN: LicenceRSN serves as the primary key and must be unique for each record. Generate an alert if it is not unique.

Business Name: When status is issued, either BusinessName or BusinessTradeName should be mentioned. Both should not be left blank.

Postal Code: The postal code must adhere to the Canadian format(LDL DLD) for records with 'Issued' status where L represents a Letter and D represents a Digit.

City: The city name must be valid and exist in the City Lookup Table for records with 'Issued' status.

Province: Province abbreviations must consist of two capitalized letters and should also match entries in the province table for records marked as 'Issued'.