British Columbia Institute of Technology

COMP 3839 DQI PROJECT

2020

DQ Profiling of Business Licences

Data Quality Improvement

Instructed by Gordon Hamilton

A01075509: Younsook Choi

Contents

Overview of Project

To operate a business in the City of Vancouver requires a valid business licence which can be obtained from the City's Licence Office and is valid for the rest of the calendar year unless stated otherwise. This dataset contains only 2019 business license records from the City of Vancouver Open Data Portal. There are about 63 thousand records with 24 columns on the dataset. The columns contain some basic information about business licences such as business name, type, licence number, address, number of employees, issued date, and expired date. There will be a small chance of low quality that is generated from data entry errors. The dataset is found from this link: City of Vancouver Open Data Portal

Column Name	Description
FOLDERYEAR	First two characters of the Business Licence Number, representing the year issued
LicenceRSN	Unique identifier for each business licence generated by the system
LicenceNumber	This field is composed with two parts: The first two digits for issued year, followed by a hyphen and a six-digit system generated number for 9 characters.
LicenceRevisionNumber	The licence version is in a two-digit format. 00 means the original version and this number increases as new revisions are created.
BusinessName	The ownership of the business
BusinessTradeName	Name under which business is usually conducted
Status	Current status of the business licence.
IssuedDate	The date when the business licence is issued and printed.
ExpiredDate	The date that the business licence expires. Most licences expire on December 31st.
BusinessType	Description of the business activity, usually in accordance with the definition in the licence By-Law No. 4450
BusiessSubType	Sub-category(s) of the main business type
Unit	Official space identifier for a building
UnitType	Description of a location other than a house or building with a simple street address where the business is located.
House	The number assigned to an address where the business is located.
Street	The name of the street where the business is located.
City	The name of the city where the business is located.
Province	The name of province or state where the business is located.
Country	The country where the business is located. Two characters
PostalCode	Postal code or zipcode
LocalArea	Local area boundary
NumberofEmployees	Number of staff employed with the business
FeePaid	Total amount of licence fee paid in Canadian dollars
ExtractDate	Date when data was extracted from source data system
Geom	Spatial representation of feature

Data Extracted from City of Vancouver Open Data Portal

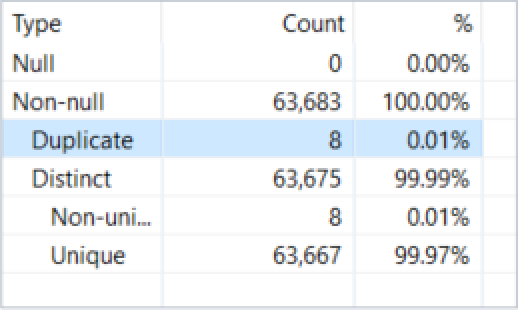
This data quality profiling project is assigned to Charles Smith, a Subject Matter Expert. He is the Licence Office Manager in Development and Building Services Centre. He will review the initial assessment that is performed by the DQ Analyst. Based on this he will give each issue found a priority, this will help to identify the most critical areas to address. He may request further research if it is necessary. After the reviewing the research done by the Data Analyst, he will suggest some DQ rules that will prevent any future data entries from causing major issues.

Initial Assessment

LicenceRSN : Duplicates

The column LicenceRSN is the primary key that identifies each row in the data set. Since there are 8 duplicated values, it is against the data integrity rules. This column value is generated by the system, so this anomaly is an important system error. Looking at the basic counts, 8 of the Licence RSN are duplicated.

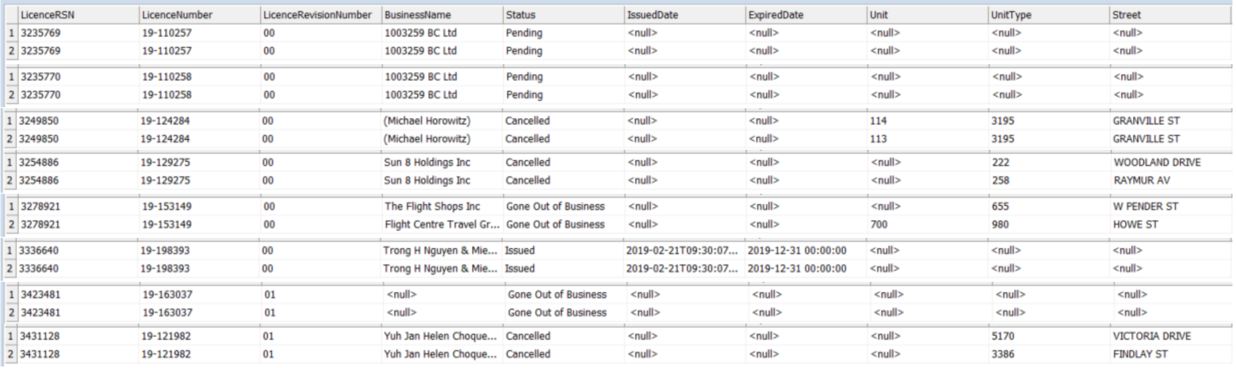
Basic Counts



Type	Count	%
Null	0	0.00%
Non-null	63,683	100.00%
Duplicate	8	0.01%
Distinct	63,675	99.99%
Non-uni...	8	0.01%
Unique	63,667	99.97%

Most of the address and issued date of those records is null and have varying license statuses. There is no such a pattern that can cause the duplicates as seen below in the table of duplicated LicenceRSN.

Duplicated LicenceRSN



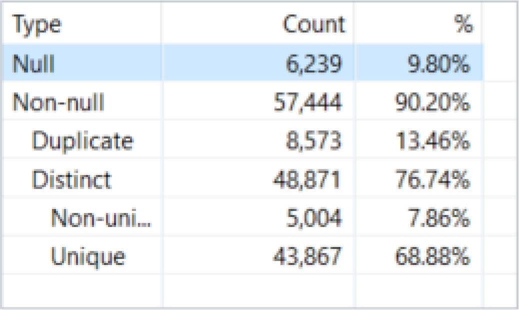
	LicenceRSN	LicenceNumber	LicenceRevisionNumber	BusinessName	Status	IssuedDate	ExpiredDate	Unit	UnitType	Street
1	3235769	19-110257	00	1003259 BC Ltd	Pending	<null>	<null>	<null>	<null>	<null>
2	3235769	19-110257	00	1003259 BC Ltd	Pending	<null>	<null>	<null>	<null>	<null>
1	3235770	19-110258	00	1003259 BC Ltd	Pending	<null>	<null>	<null>	<null>	<null>
2	3235770	19-110258	00	1003259 BC Ltd	Pending	<null>	<null>	<null>	<null>	<null>
1	3249850	19-124284	00	(Michael Horowitz)	Cancelled	<null>	<null>	114	3195	GRANVILLE ST
2	3249850	19-124284	00	(Michael Horowitz)	Cancelled	<null>	<null>	113	3195	GRANVILLE ST
1	3254886	19-129275	00	Sun 8 Holdings Inc	Cancelled	<null>	<null>	<null>	222	WOODLAND DRIVE
2	3254886	19-129275	00	Sun 8 Holdings Inc	Cancelled	<null>	<null>	<null>	258	RAYMUR AV
1	3278921	19-153149	00	The Flight Shops Inc	Gone Out of Business	<null>	<null>	<null>	655	W PENDER ST
2	3278921	19-153149	00	Flight Centre Travel Gr...	Gone Out of Business	<null>	<null>	700	980	HOWE ST
1	3336640	19-198393	00	Trong H Nguyen & Mie...	Issued	2019-02-21T09:30:07...	2019-12-31 00:00:00	<null>	<null>	<null>
2	3336640	19-198393	00	Trong H Nguyen & Mie...	Issued	2019-02-21T09:30:07...	2019-12-31 00:00:00	<null>	<null>	<null>
1	3423481	19-163037	01	<null>	Gone Out of Business	<null>	<null>	<null>	<null>	<null>
2	3423481	19-163037	01	<null>	Gone Out of Business	<null>	<null>	<null>	<null>	<null>
1	3431128	19-121982	01	Yuh Jan Helen Choque...	Cancelled	<null>	<null>	<null>	5170	VICTORIA DRIVE
2	3431128	19-121982	01	Yuh Jan Helen Choque...	Cancelled	<null>	<null>	<null>	3386	FINDLAY ST

BusinessName: NULL values

The first thing of the profile on Business Name that stands out is about 10% of the records are empty so we cannot expect to rely on this field heavily.

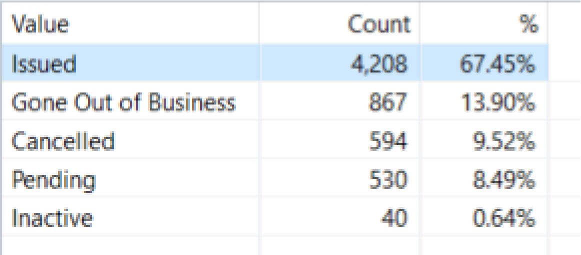
Looking at the below Frequency Analysis, over 4,200 business licences are issued, and all have BusinessTradeName of NULL in Short-Term Rental firms. Short-Term Rental business types do not have both BusinessNames and BusinessTradeNames. Only 2 records have BusinessTradeNames in other business type – Educational, Beauty Services.

Basic Counts



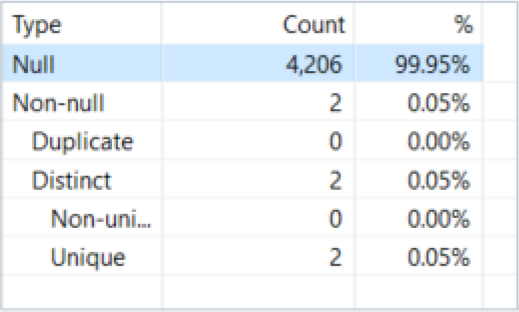
Type	Count	%
Null	6,239	9.80%
Non-null	57,444	90.20%
Duplicate	8,573	13.46%
Distinct	48,871	76.74%
Non-uni...	5,004	7.86%
Unique	43,867	68.88%

Frequency Analysis of Status when BusinessName=NULL



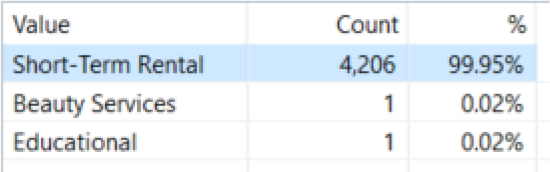
Value	Count	%
Issued	4,208	67.45%
Gone Out of Business	867	13.90%
Cancelled	594	9.52%
Pending	530	8.49%
Inactive	40	0.64%

Basic Counts of BusinessTradeName when Status=Issued and BusinessName=NULL



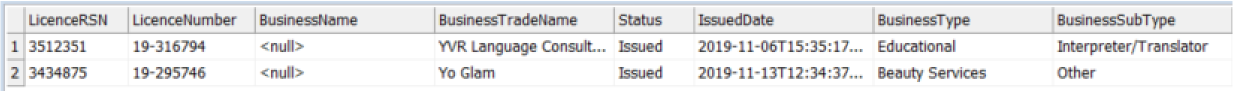
Type	Count	%
Null	4,206	99.95%
Non-null	2	0.05%
Duplicate	0	0.00%
Distinct	2	0.05%
Non-uni...	0	0.00%
Unique	2	0.05%

Frequency Analysis of BusinessType when Status=Issued and BusinessName=NULL



Value	Count	%
Short-Term Rental	4,206	99.95%
Beauty Services	1	0.02%
Educational	1	0.02%

Drill-Through: BusinessTradeName <> NULL when Status=Issued and BusinessName=NULL

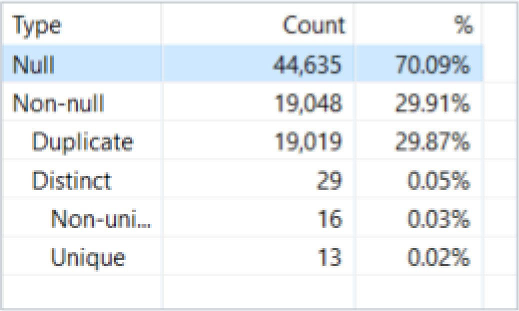


	LicenceRSN	LicenceNumber	BusinessName	BusinessTradeName	Status	IssuedDate	BusinessType	BusinessSubType
1	3512351	19-316794	<null>	YVR Language Consult...	Issued	2019-11-06T15:35:17...	Educational	Interpreter/Translator
2	3434875	19-295746	<null>	Yo Glam	Issued	2019-11-13T12:34:37...	Beauty Services	Other

UnitType: Various values for a type

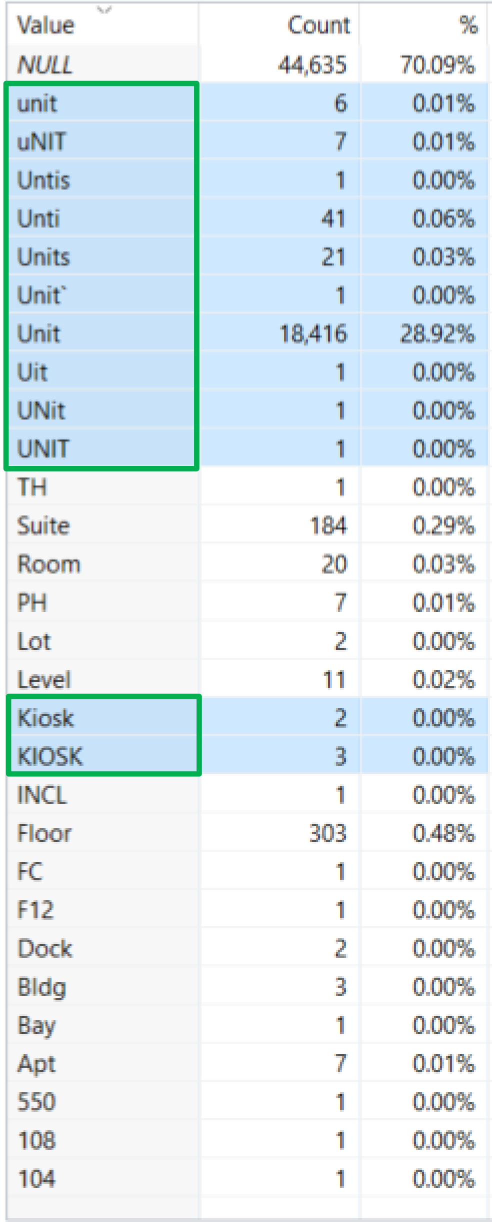
In the UnitType column, there are multiple formats for a unit type including typo, plural format, and case sensitive. Looking at Frequency Analysis, type “Unit” has 10 variations and type “Kiosk” has 2 variations for instance.

Basic Counts



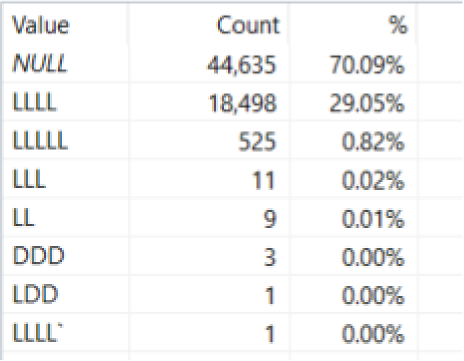
Type	Count	%
Null	44,635	70.09%
Non-null	19,048	29.91%
Duplicate	19,019	29.87%
Distinct	29	0.05%
Non-uni...	16	0.03%
Unique	13	0.02%

Frequency Analysis



Value	Count	%
NULL	44,635	70.09%
unit	6	0.01%
uNIT	7	0.01%
Untis	1	0.00%
Unti	41	0.06%
Units	21	0.03%
Unit`	1	0.00%
Unit	18,416	28.92%
Uit	1	0.00%
UNit	1	0.00%
UNIT	1	0.00%
TH	1	0.00%
Suite	184	0.29%
Room	20	0.03%
PH	7	0.01%
Lot	2	0.00%
Level	11	0.02%
Kiosk	2	0.00%
KIOSK	3	0.00%
INCL	1	0.00%
Floor	303	0.48%
FC	1	0.00%
F12	1	0.00%
Dock	2	0.00%
Bldg	3	0.00%
Bay	1	0.00%
Apt	7	0.01%
550	1	0.00%
108	1	0.00%
104	1	0.00%

Mask Analysis

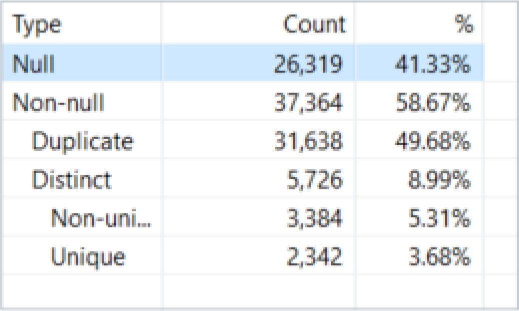


Value	Count	%
NULL	44,635	70.09%
LLLL	18,498	29.05%
LLLLL	525	0.82%
LLL	11	0.02%
LL	9	0.01%
DDD	3	0.00%
LDD	1	0.00%
LLLL`	1	0.00%

PostalCode: NULL values

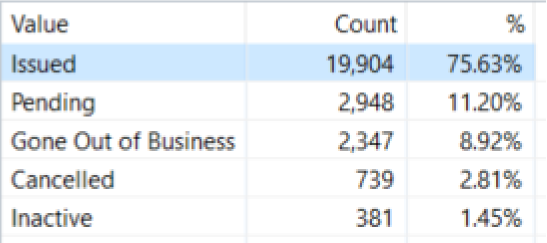
On the information of basic counts, it is noticeable that over 40% of postal codes are empty and about 76% of those licences are issued. It will cost for post delivery with wrong postal code.

Basic Counts



Type	Count	%
Null	26,319	41.33%
Non-null	37,364	58.67%
Duplicate	31,638	49.68%
Distinct	5,726	8.99%
Non-uni...	3,384	5.31%
Unique	2,342	3.68%

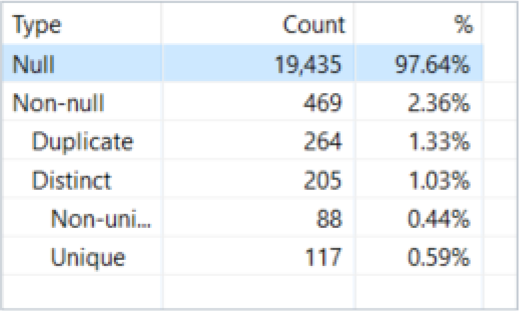
Frequency Analysis of Status when PostalCode=NULL



Value	Count	%
Issued	19,904	75.63%
Pending	2,948	11.20%
Gone Out of Business	2,347	8.92%
Cancelled	739	2.81%
Inactive	381	1.45%

Lookngi at the basic counts below, there are about 98% of rows that have NULL value on both Street and PostalCode. It may cause significant delays when mailing licences. Is there any additional data that would help fix the issues with the postal codes?

Basic Counts of Street when PostalCode=NULL and Status=Issued

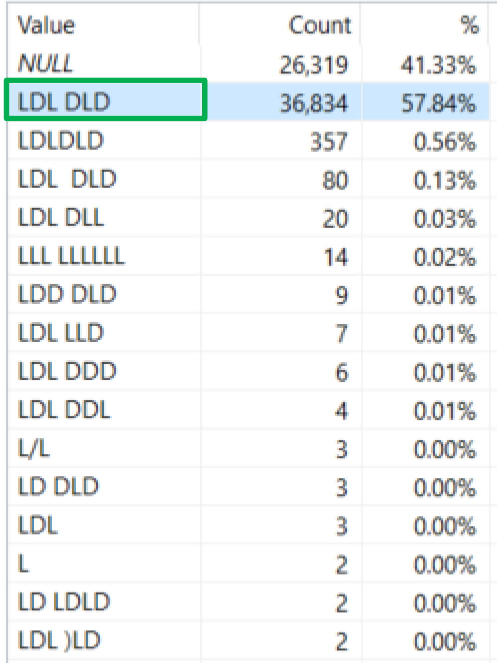


Type	Count	%
Null	19,435	97.64%
Non-null	469	2.36%
Duplicate	264	1.33%
Distinct	205	1.03%
Non-uni...	88	0.44%
Unique	117	0.59%

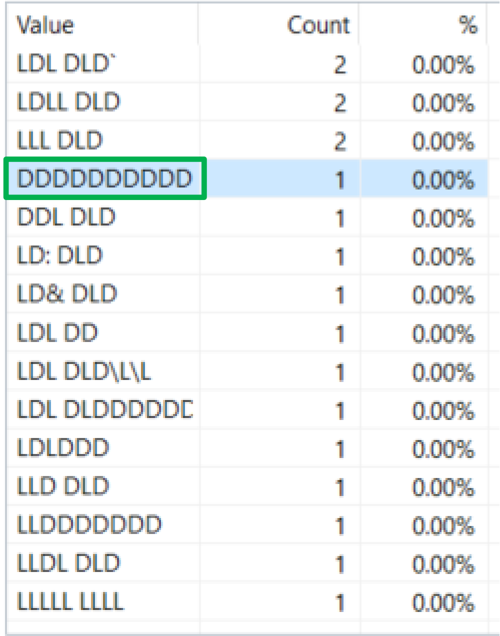
PostalCode: Various formats

There are many postal code formats, and we can assume that ‘LDL DLD’ is the proper format since about 60% of records are in that format. However, there might be another variation ‘DDDDD’ for the American standard for postal codes.

Mask Analysis



Value	Count	%
NULL	26,319	41.33%
LDL DLD	36,834	57.84%
LDLDLD	357	0.56%
LDL DLD	80	0.13%
LDL DLL	20	0.03%
LLL LLLLLL	14	0.02%
LDD DLD	9	0.01%
LDL LLD	7	0.01%
LDL DDD	6	0.01%
LDL DDL	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%

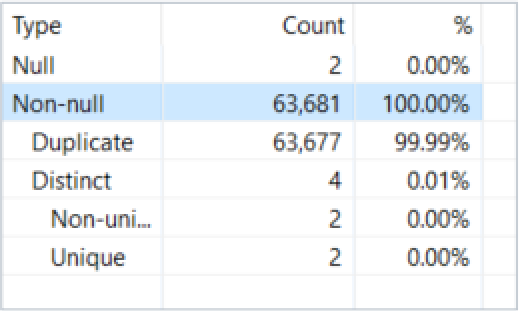


Value	Count	%
LDL DLD`	2	0.00%
LDLL DLD	2	0.00%
LLL DLD	2	0.00%
DDDDDDDDDD	1	0.00%
DDL DLD	1	0.00%
LD: DLD	1	0.00%
LD& DLD	1	0.00%
LDL DD	1	0.00%
LDL DLD\L\L	1	0.00%
LDL DLDDDDDD	1	0.00%
LDLDDD	1	0.00%
LLD DLD	1	0.00%
LLDDDDDDD	1	0.00%
LLDL DLD	1	0.00%
LLLLL LLLL	1	0.00%

Province: Invalid values

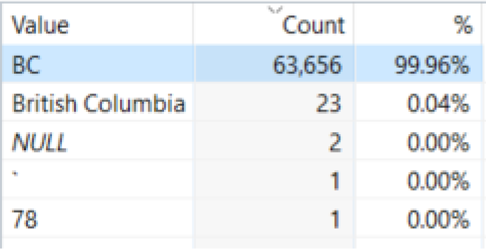
There are several records with the full province name while over 99% of records use the abbreviation. A couple of records have NULL value on the province column. As well there are a couple of records that have an invalid value in the province column.

Basic Counts



Type	Count	%
Null	2	0.00%
Non-null	63,681	100.00%
Duplicate	63,677	99.99%
Distinct	4	0.01%
Non-uni...	2	0.00%
Unique	2	0.00%

Frequency Analysis



Value	Count	%
BC	63,656	99.96%
British Columbia	23	0.04%
NULL	2	0.00%
`	1	0.00%
78	1	0.00%

SME Review of Initial Assessment

LicenceRSN : Duplicates (High Priority)

This duplicate issue should take the top priority for the data quality improvement process since it is against the integrity data rules and it is generated by the system. To find out any patterns on the duplicate problem, it is better to analyze the volume of issued licences by month. It might show erratic characteristics.

PostalCode: Invalid Values (Medium Priority)

This problem has a high business impact that costs post delivery with wrong postal codes. Solving this issue should reduce costs and time associated with correcting the mailing information. For PostalCode values of Null, this issue can be solved by determining Postal Code based on address so we can fill in these fields later. For the case of missing both Street and PostalCode, the information from business accounts might be used for determining the missing information.

UnitType: Various values for a type (Medium Priority)

This variation issue could also cause additional post delivery costs. As with the Postal Code issue, fixing this will save time and postal costs. This variation issue could be removed by providing the list of unit type at user entry.

BusinessName: NULL values (Low Priority)

This issue is considered a minor data quality issue since BusinessName for Short-Term Business licences is not mandatory, so the data quality is fine. All other types of business licences have business names so that is the reason this is minor issue for now. We can determine Business Name based on Licence Number so we can fill in many of these fields later based on other information.

Province: Invalid values (Low Priority)

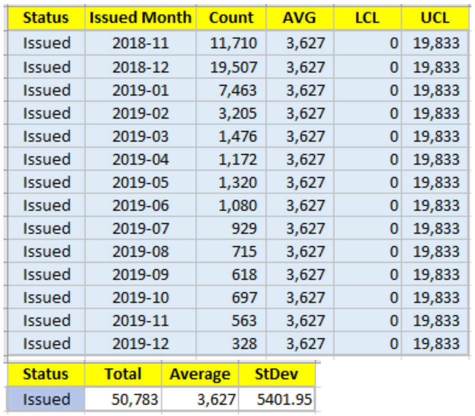
The value of province has the least business impact among the other issues so this matter can be processed as the last of DQI process. The invalid province value can be determined by Postal Code.

Further Research for the SME

The volume of issued licences per month

For statistical process control analysis, all issued licences are counted by month of issuedDate. An unusual pattern is found from the analysis. Looking at SPC Chart, there are extremely high volumes in November and December 2018, while those same month in 2019 stays relatively low. Is there any predictor or pattern for it?

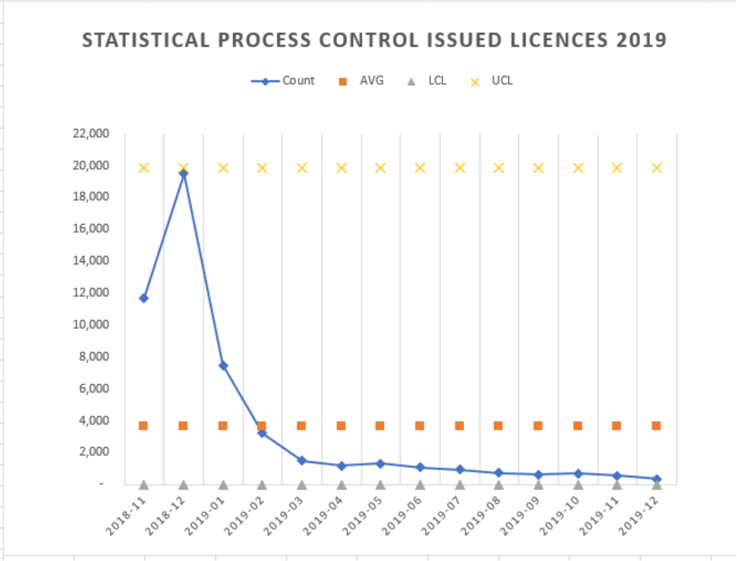
SPC Analysis



Status	Issued Month	Count	AVG	LCL	UCL
Issued	2018-11	11,710	3,627	0	19,833
Issued	2018-12	19,507	3,627	0	19,833
Issued	2019-01	7,463	3,627	0	19,833
Issued	2019-02	3,205	3,627	0	19,833
Issued	2019-03	1,476	3,627	0	19,833
Issued	2019-04	1,172	3,627	0	19,833
Issued	2019-05	1,320	3,627	0	19,833
Issued	2019-06	1,080	3,627	0	19,833
Issued	2019-07	929	3,627	0	19,833
Issued	2019-08	715	3,627	0	19,833
Issued	2019-09	618	3,627	0	19,833
Issued	2019-10	697	3,627	0	19,833
Issued	2019-11	563	3,627	0	19,833
Issued	2019-12	328	3,627	0	19,833

Status	Total	Average	StDev
Issued	50,783	3,627	5401.95

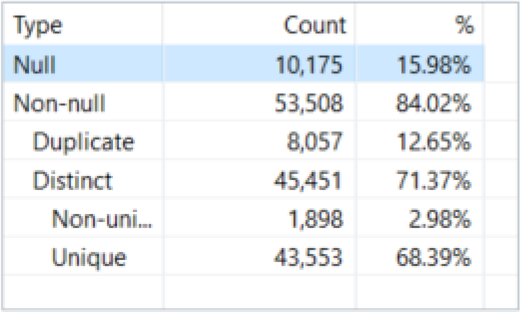
SPC Chart



Invalid IssuedDate

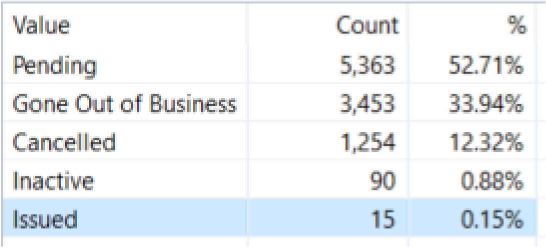
It is found during analyzing the volume of issued licences by month of IssuedDate. There are several records that have an invalid date or null value on issued business licences. 37 dates from issued records are out of range of date which is after expired date, and 15 issued licences do not have issued date. These anomalies might be from data entry errors.

Basic Counts



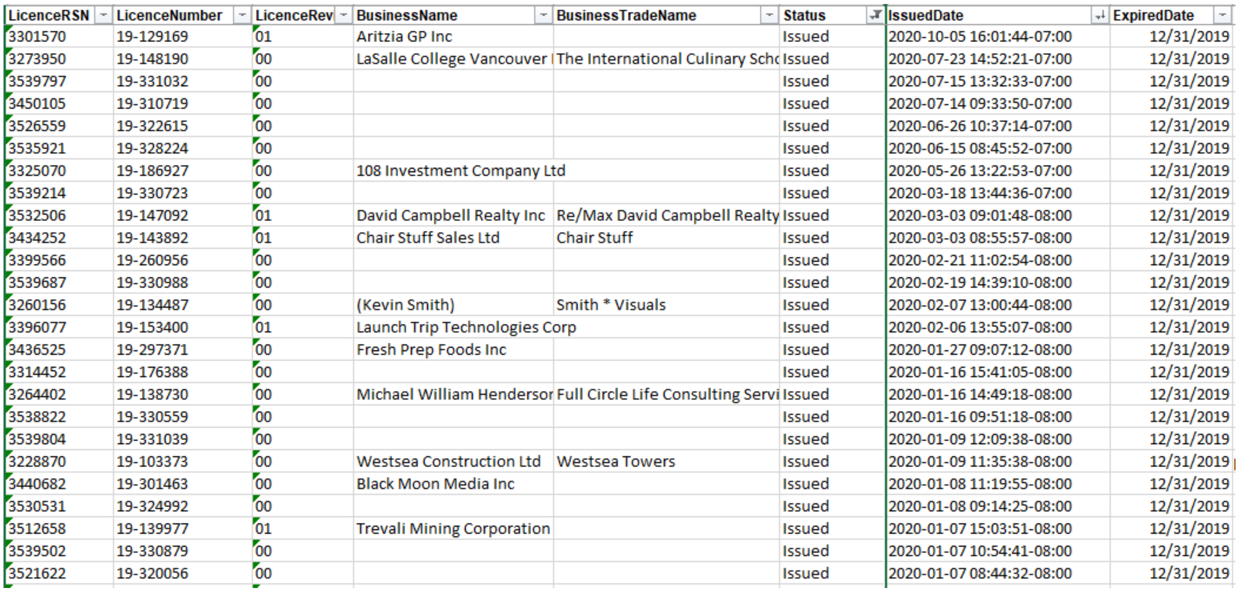
Type	Count	%
Null	10,175	15.98%
Non-null	53,508	84.02%
Duplicate	8,057	12.65%
Distinct	45,451	71.37%
Non-uni...	1,898	2.98%
Unique	43,553	68.39%

Frequency Analysis of Status when IssuedDate=Null



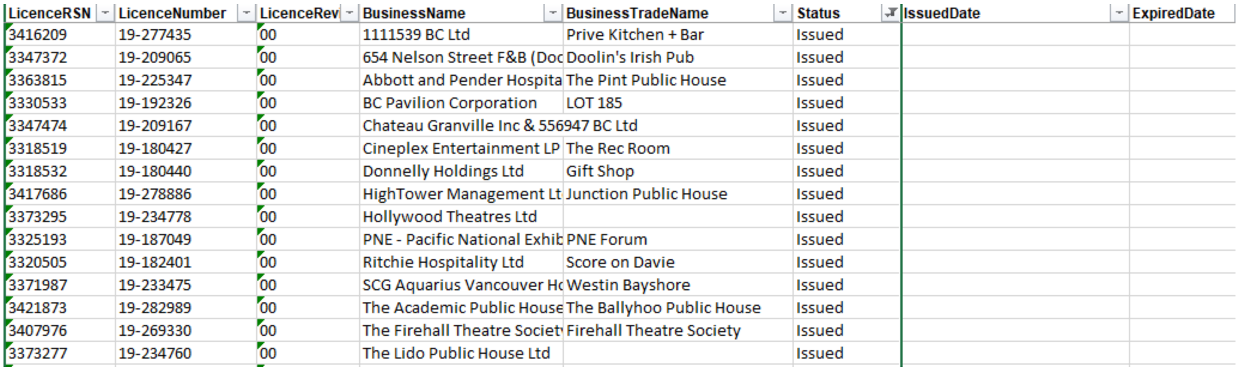
Value	Count	%
Pending	5,363	52.71%
Gone Out of Business	3,453	33.94%
Cancelled	1,254	12.32%
Inactive	90	0.88%
Issued	15	0.15%

Issued date after expired date



LicenceRSN	LicenceNumber	LicenceRev	BusinessName	BusinessTradeName	Status	IssuedDate	ExpiredDate
3301570	19-129169	01	Aritzia GP Inc		Issued	2020-10-05 16:01:44-07:00	12/31/2019
3273950	19-148190	00	LaSalle College Vancouver I	The International Culinary Schc	Issued	2020-07-23 14:52:21-07:00	12/31/2019
3539797	19-331032	00			Issued	2020-07-15 13:32:33-07:00	12/31/2019
3450105	19-310719	00			Issued	2020-07-14 09:33:50-07:00	12/31/2019
3526559	19-322615	00			Issued	2020-06-26 10:37:14-07:00	12/31/2019
3535921	19-328224	00			Issued	2020-06-15 08:45:52-07:00	12/31/2019
3325070	19-186927	00	108 Investment Company Ltd		Issued	2020-05-26 13:22:53-07:00	12/31/2019
3539214	19-330723	00			Issued	2020-03-18 13:44:36-07:00	12/31/2019
3532506	19-147092	01	David Campbell Realty Inc	Re/Max David Campbell Realty	Issued	2020-03-03 09:01:48-08:00	12/31/2019
3434252	19-143892	01	Chair Stuff Sales Ltd	Chair Stuff	Issued	2020-03-03 08:55:57-08:00	12/31/2019
3399566	19-260956	00			Issued	2020-02-21 11:02:54-08:00	12/31/2019
3539687	19-330988	00			Issued	2020-02-19 14:39:10-08:00	12/31/2019
3260156	19-134487	00	(Kevin Smith)	Smith * Visuals	Issued	2020-02-07 13:00:44-08:00	12/31/2019
3396077	19-153400	01	Launch Trip Technologies Corp		Issued	2020-02-06 13:55:07-08:00	12/31/2019
3436525	19-297371	00	Fresh Prep Foods Inc		Issued	2020-01-27 09:07:12-08:00	12/31/2019
3314452	19-176388	00			Issued	2020-01-16 15:41:05-08:00	12/31/2019
3264402	19-138730	00	Michael William Hendersor	Full Circle Life Consulting Servi	Issued	2020-01-16 14:49:18-08:00	12/31/2019
3538822	19-330559	00			Issued	2020-01-16 09:51:18-08:00	12/31/2019
3539804	19-331039	00			Issued	2020-01-09 12:09:38-08:00	12/31/2019
3228870	19-103373	00	Westsea Construction Ltd	Westsea Towers	Issued	2020-01-09 11:35:38-08:00	12/31/2019
3440682	19-301463	00	Black Moon Media Inc		Issued	2020-01-08 11:19:55-08:00	12/31/2019
3530531	19-324992	00			Issued	2020-01-08 09:14:25-08:00	12/31/2019
3512658	19-139977	01	Trevali Mining Corporation		Issued	2020-01-07 15:03:51-08:00	12/31/2019
3539502	19-330879	00			Issued	2020-01-07 10:54:41-08:00	12/31/2019
3521622	19-320056	00			Issued	2020-01-07 08:44:32-08:00	12/31/2019

NULL Issued date on Issued Business Licences



LicenceRSN	LicenceNumber	LicenceRev	BusinessName	BusinessTradeName	Status	IssuedDate	ExpiredDate
3416209	19-277435	00	1111539 BC Ltd	Prive Kitchen + Bar	Issued		
3347372	19-209065	00	654 Nelson Street F&B (Doc	Doolin's Irish Pub	Issued		
3363815	19-225347	00	Abbott and Pender Hospita	The Pint Public House	Issued		
3330533	19-192326	00	BC Pavilion Corporation	LOT 185	Issued		
3347474	19-209167	00	Chateau Granville Inc & 556947 BC Ltd		Issued		
3318519	19-180427	00	Cineplex Entertainment LP	The Rec Room	Issued		
3318532	19-180440	00	Donnelly Holdings Ltd	Gift Shop	Issued		
3417686	19-278886	00	HighTower Management Lt	Junction Public House	Issued		
3373295	19-234778	00	Hollywood Theatres Ltd		Issued		
3325193	19-187049	00	PNE - Pacific National Exhib	PNE Forum	Issued		
3320505	19-182401	00	Ritchie Hospitality Ltd	Score on Davie	Issued		
3371987	19-233475	00	SCG Aquarius Vancouver Hc	Westin Bayshore	Issued		
3421873	19-282989	00	The Academic Public House	The Ballyhoo Public House	Issued		
3407976	19-269330	00	The Firehall Theatre Societ	Firehall Theatre Society	Issued		
3373277	19-234760	00	The Lido Public House Ltd		Issued		

SME Review of Further Research

The statistical process control on issued licences is under control, and the high volume in November and December 2018 was just a temporary occurrence.

It is good that the anomaly of invalid issued date is found. It is very odd to have an issued date after the expired date or a null value. If the status of licence is issued and the issued date is null, or if the issued date is greater than expired date, then it will be considered an invalid value. These constraints should be implemented in the system.

SME Suggests Some DQ Rules

LicenceRSN

LicenceRSN should not be duplicated since every row in the data set uniquely identifies by LicenceRSN. The system must detect the duplicate issue when saving a new data of business licence.

BusinessName

BusinessName column is required when the Status is Issued and the BusinessType is not a Short-Term Rental. Short-Term Rental is a temporary accommodation business, so it is not necessary to enter BusinessName.

UnitType

The unit type describes a location where the business is located. The unit type may have values of unit, suite, room, building, apartment, etc. A value must be picked when entering a new licence from the list of choices. It would help to also have a value for a unit type that falls outside this category but is common, such as a residential street address. A blank is considered an invalid value when the Status is Issued.

PostalCode

Postal code is related to country. It is good to validate the format when entering a new postal code based on the value of country.

Province

The province column is related to country column. It is good to provide the list of provinces with abbreviation and full names based on the value of country. The official recognised abbreviations for provinces should be the only options.

IssuedDate

Issued date may not be blank for issued licences and it should be before expiry date. A date must be picked when entering a new licence and thus the constraint of before the expiry date should be implemented at the data entry level. If issued date is not before the expiry date, it will be considered an invalid value.