VANCOUVER BUSINESS LICENCES – DATA QUALITY ANALYSIS

Tanvir Bajwa

Contents

Overview of Project

The data that I have decided to analyze and ultimately improve the quality is year 2018 of the Vancouver business licence database found here: https://opendata.vancouver.ca/explore/dataset/business-licences/table/?disjunctive.status&disjunctive.businesssubtype&refine.issueddate=2018.

The data consists of 58,628 records and 24 columns describing it.

The project will start off with an initial assessment of the data (column by column) and work in conjunction the person responsible for data input to understand and fix the issues. The project aims to identify the deficiencies in the current database set up and provide some recommendations to maintains data integrity going forward.

Definitions

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	9-character field where first two digits are the year issued followed by a six-digit system generated number separated by a hyphen.
LicenceRevisionNumber	2-digit field representing document version. 00 indicates licence is unrevised.
BusinessName	Name under which a business is registered.
BusinessTradeName	Name under which a business operates (sometimes different than registered name)
Status	Shows status of business licence – possible values are 'cancelled', 'Gone Out of Business', 'Inactive', 'Issued', 'Pending'.
IssuedDate	Date when licence issued and printed.
ExpiredDate	Date when licence expires.
BusinessType	Description of business activity as per By-Law No. 4450.
BusinessSubType	Sub-category of business type (if any).
Unit	Alpha numeric field indicating space in a building.
UnitType	Description of location other than house or building i.e., Suite, Apartment etc.
House	Number assigned to an address where the business is located.
Street	Name of street where business is located.
City	Name of the city where business is located.
Province	Name of the province/state where business is located.

Column Name	Description
Country	Two-character field that signifies name of country where the business is located.
PostalCode	Series of letter and/or digits attached to the address of the business.
LocalArea	Also known as local planning areas. Vancouver has 22 local areas.
NumberofEmployees	Number of staff employed with the business.
ExtractDate	Date when data was extracted from the source data system.
Geom	Spatial representation of feature.

Initial Assessment

LicenceRSN: Duplicates

As per the definition, all values in this column should be unique and therefore can be used as a primary key to identify records. Referential integrity of the entire database is in question as there are two non-unique records as shown below.

Type	Count	%
Null	0	0.00%
Non-null	58,628	100.00...
Duplicate	2	0.00%
Distinct	58,626	100.00...
Non-unique	2	0.00%
Unique	58,624	99.99%

In both the duplication instances, the issued date, status, revision number, and expiration date are the same. In the case of licence # 3178764, the business name column has different values. For licence #3012672, the address recorded is different.

FOI	LicenceRSN	LicenceNum	Licen	BusinessName	Status	IssuedDate	ExpiredDate	BusinessType	BusinessSubT	Unit	UnitType	House	Street	City
18	3178764	18-594011	0	(Keiko Yokoyama)	Issued	2018-07-19 11:09	2018-12-31	Apartment House Strata		607	Unit	1068	HORNBY ST	Vancouver
18	3178764	18-594011	0	Iemitsu Yokoyama & Keiko Yokoyama	Issued	2018-07-19 11:09	2018-12-31	Apartment House Strata		607	Unit	1068	HORNBY ST	Vancouver
18	3012672	18-444398	0	Vancouver City Savings Credit Union	Issued	2018-07-24 13:55	2018-12-31	Financial Services	Bank Machine			1610	ROBSON ST	Vancouver
18	3012672	18-444398	0	Vancouver City Savings Credit Union	Issued	2018-07-24 13:55	2018-12-31	Financial Services	Bank Machine			183	TERMINAL AV	Vancouver

PostalCode: Invalid records

As per the database, it is evident that there are records from Canada and the US; therefore, it makes sense to have two formats. However, as seen below, there are many different formats to record the postal code which indicates lack of consistency in this column. This causes issues such as returned mail which impacts the bottom line as it is ultimately wastage of resources.

Mask Analysis

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

Value	Count	%
LDL DLD	32,523	55.47%
NULL	25,624	43.71%
LDLDLD	315	0.54%
LDL DLD	77	0.13%
LDL DLL	16	0.03%
LLL LLLLLL	10	0.02%
LDL DDD	8	0.01%
DDDDD	6	0.01%
LDD DLD	6	0.01%
LDL LLD	5	0.01%
LD DLD	4	0.01%
LDL	4	0.01%
LDL DDL	4	0.01%
LDL)LD	3	0.01%
LLL DLD	3	0.01%
L	2	0.00%
LDL DLD`	2	0.00%
LDLL DLD	2	0.00%
LLD DLD	2	0.00%
DDDDDD	1	0.00%
DDL DLD	1	0.00%
LD DLL	1	0.00%
LD LDLD	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 DLDDDDI	1	0.00%
LDLDDD	1	0.00%
LL	1	0.00%
LLDDDDDDD	1	0.00%

PostalCode: NULL values

There is a large amount of missing data in this column as shown below. The value in this column would be required to mail out key information to licence holders. Of the records where postal code is null, 24,747 or 96.57% show status as issued. The absence of a value in this column could have similar implications on the bottom line due to possible mail return. There appears to be lack of consistency in the column as well.

Type	Count	%
Null	25,624	43.71%
Non-null	33,004	56.29%
Duplicate	27,821	47.45%
Distinct	5,183	8.84%
Non-...	3,311	5.65%
Unique	1,872	3.19%

PostalCode	
Row Labels	**Count of Status**
Cancelled	14
Gone Out of Business	333
Inactive	462
Issued	24747
Pending	68
Grand Total	**25624**

FeePaid: NULL values

Missing data in the column could lead to incorrect reporting on profitability as the fees collected is a source of revenue. The overall count of null records is minimal and maximum nulls recorded under status = issued.

Type	Count	%
Null	172	0.29%
Non-null	58,456	99.71%
Duplicate	56,821	96.92%
Distinct	1,635	2.79%
Non-...	832	1.42%
Unique	803	1.37%

FeePaid	(blank)
Row Labels	**Count of Status**
Cancelled	6
Gone Out of Business	13
Inactive	4
Issued	145
Pending	4
Grand Total	**172**

BusinessName: NULL values

8.45% of the overall records have a null value in the business name column. Out of these records, most of them have status = issued or they are active businesses. Further to the matter, most of the records with no business name or business trade name belong to the short-term rental business type. Without a business name the context is missing in the dataset – we know the ID of the record but not what business entity that record belongs to.

Type	Count	%
Null	4,955	8.45%
Non-null	53,673	91.55%
Duplicate	15,030	25.64%
Distinct	38,643	65.91%
Non-...	10,745	18.33%
Unique	27,898	47.58%

BusinessName	
BusinessTradeName	
Row Labels	**Count of Status**
Cancelled	7
Gone Out of Business	215
Inactive	36
Issued	4636
Pending	61
Grand Total	**4955**

BusinessName	
BusinessTradeName	
Status	Issued
Row Labels	**Count of Status**
Electrical Contractor	1
One-Family Dwelling	1
Scavenging	1
Short-Term Rental	4633
Grand Total	**4636**

SME Review of Initial Assessment

LicenceRSN: Duplicates (High)

This is high priority as the LicenceRSN field is used as a primary key for the entire database. In the instances shown above, it is either the address or business name that are different among the duplicates. In both the instances, one of the records is incorrect. Further analysis can be conducted to check if volume of records per month has something to do with duplication.

PostalCode: Invalid records (Medium)

This is a medium priority issue as the percentage of such records compared to the overall database is quite small; however, this issue should be tackled as soon as some larger issues are dealt with, as this affects the validity of our data. A database of correct postal codes in the proper format can be used against the Vancouver business licence database to determine the inconsistencies.

PostalCode: NULL values (High)

This is high priority, since of the records that are missing postal code, ~96% of the records are status = issued or active. This has a wide impact and further analysis to understand the issue in detail should be conducted. We need to maintain complete addresses to mail out key information for records where status = issued.

FeePaid: NULL values (Low)

This is low priority as there are only 0.29% of the records that display this anomaly. However, incorrect recordings in this column could affect regulatory/internal reporting and therefore must be dealt with. Further review into the matter based on other variables should be conducted to ascertain any trends/patterns, if any.

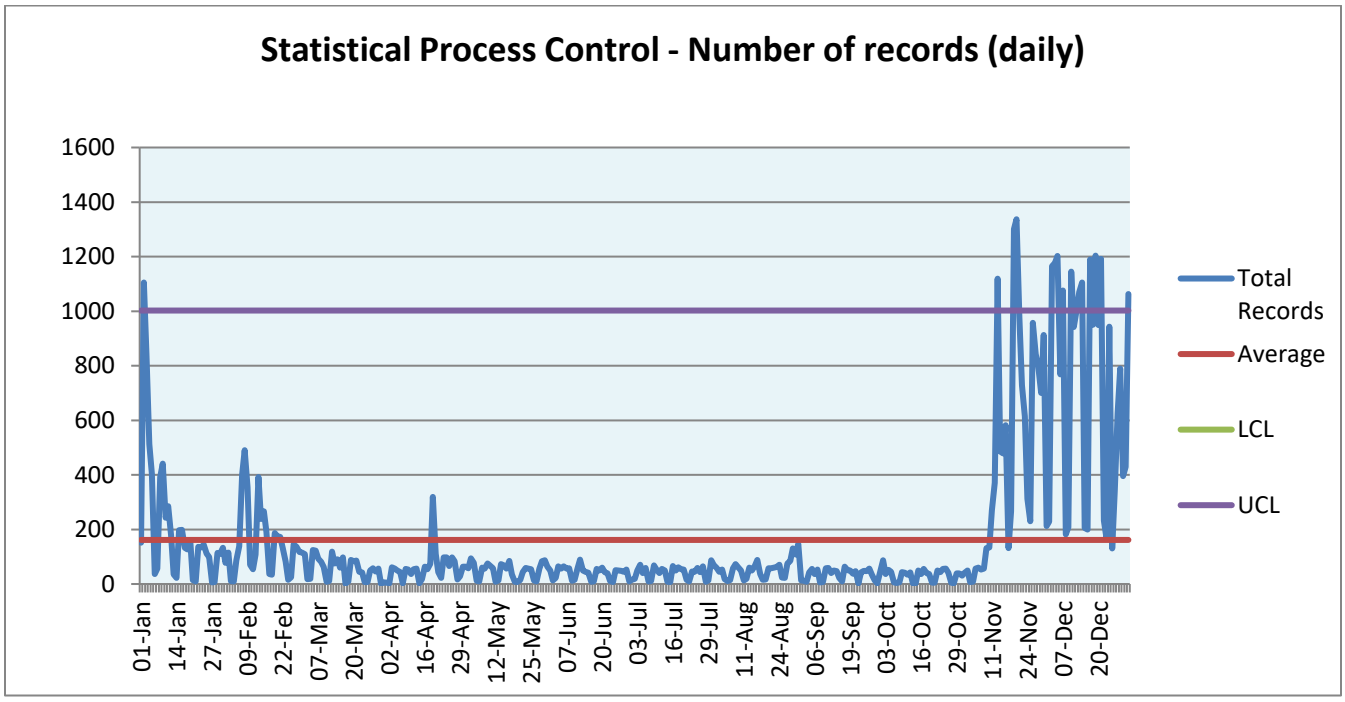
BusinessName: NULL values (High)

All registered businesses should have a name under which they operate, and we should have that recorded. Of the records that have a null value for the business name column, most of them are active which makes this a high priority issue. Further investigation can be conducted based on business type, revision number to drill down further into the issue.

Further research for SME

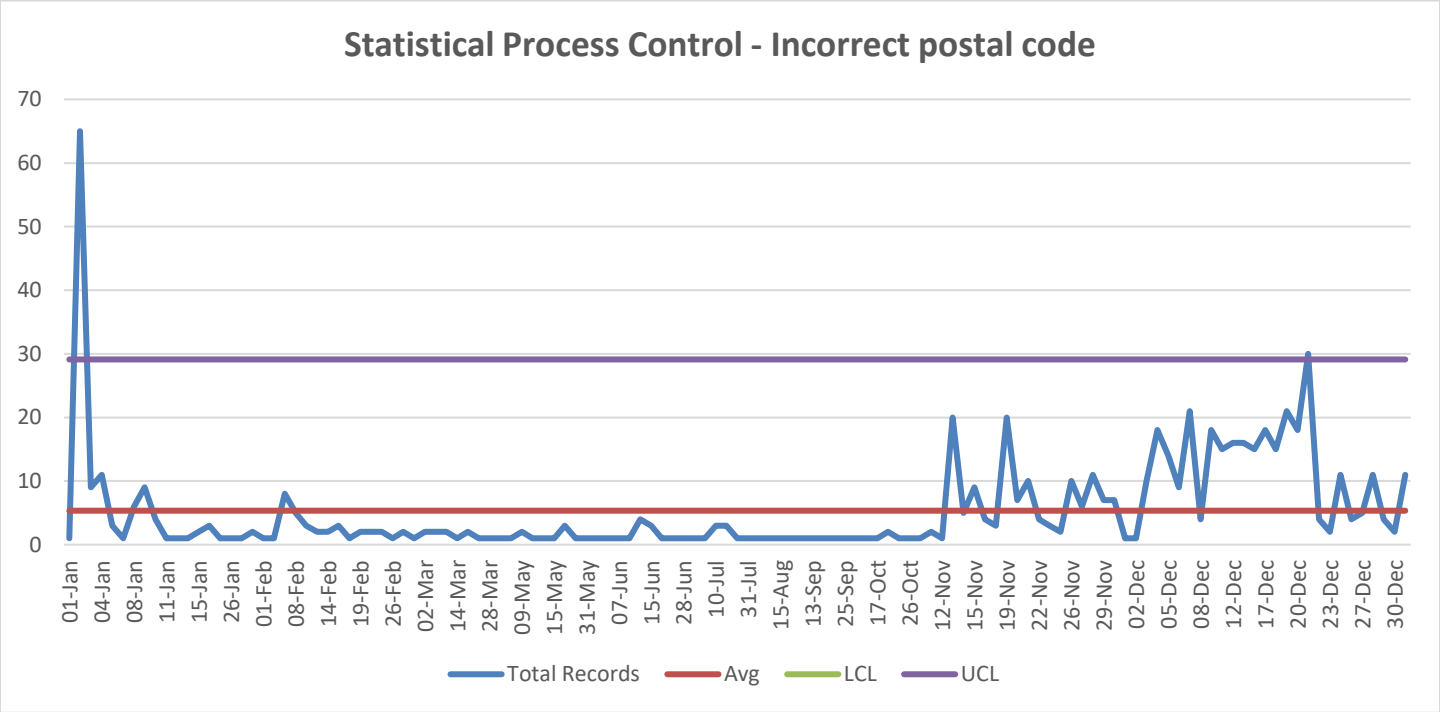
LicenceRSN: Duplicates

Further to the SME request, a statistical process control chart has been created to check whether the duplicate entries are due to an excessive number of records that were processed on a particular date. The number of records being processed daily is well within control during the month of July 2018 – the month where both the duplication instances are from. However, the number of records being processed itself breaches the upper bounds in January, November, and December 2018.



PostalCode: Invalid records

An up-to-date dataset in the correct format for Canadian and US postal values was downloaded and used to determine the incorrect entries in our database. Below is the process control chart indicating that the input for incorrect formats/values was extremely out of control in early January 2018 and slightly out of control towards the end of the year. As seen further below, most of these records show that the business is active. Determining incorrect values this way reveals other issues such as wrong entries as well.



Mask Analysis

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

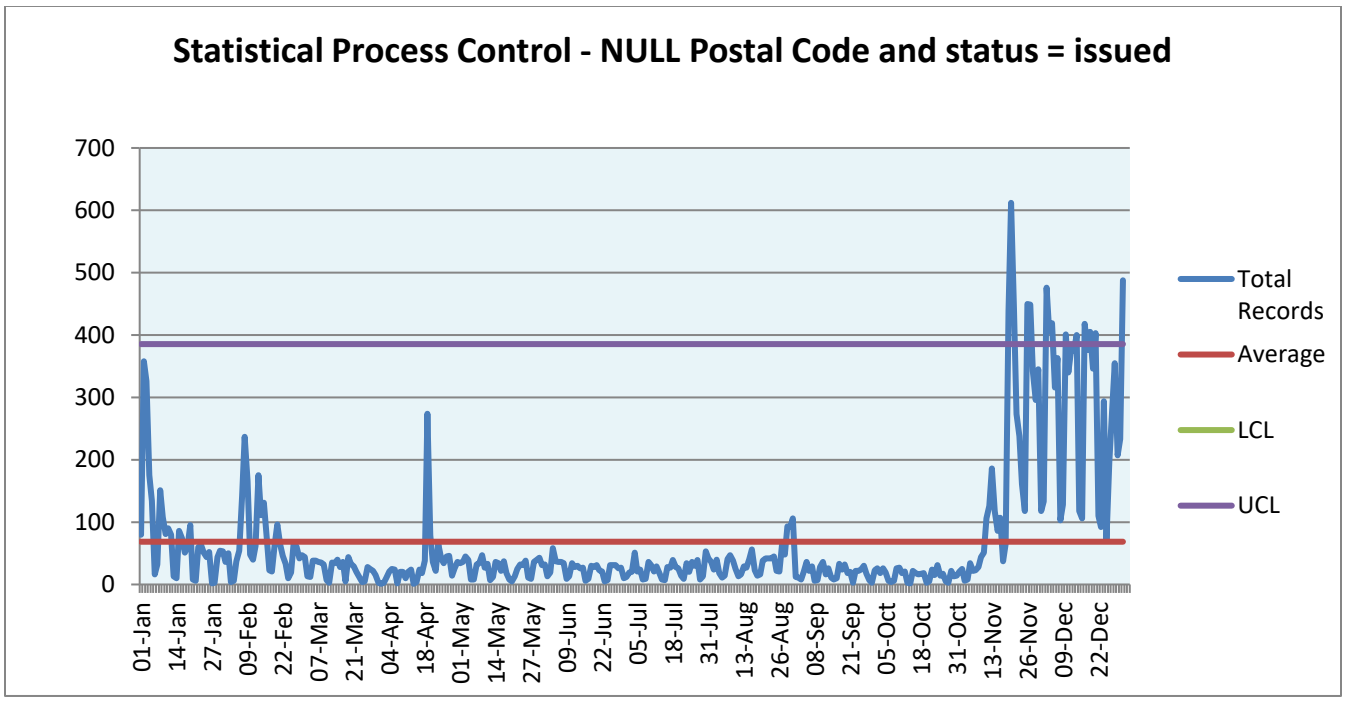
Value	Count	%
LDL DLD	32,523	55.47%
NULL	25,624	43.71%
LDLDLD	315	0.54%
LDL DLD	77	0.13%
LDL DLL	16	0.03%
LLL LLLLLL	10	0.02%
LDL DDD	8	0.01%
DDDDD	6	0.01%
LDD DLD	6	0.01%
LDL LLD	5	0.01%
LD DLD	4	0.01%
LDL	4	0.01%
LDL DDL	4	0.01%
LDL)LD	3	0.01%
LLL DLD	3	0.01%
L	2	0.00%
LDL DLD`	2	0.00%
LDLL DLD	2	0.00%
LLD DLD	2	0.00%
DDDDDD	1	0.00%
DDL DLD	1	0.00%
LD DLL	1	0.00%
LD LDLD	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 DLDDDDI	1	0.00%
LDLDDD	1	0.00%
LL	1	0.00%
LLDDDDDDD	1	0.00%

The values in the database have been measured against a current database of postal codes which have revealed the following results. The **total count of non-null incorrect entries based on format and validity** of the record as per the pivot is **higher than the number of non-null incorrect formatted records**. This indicates that there may be invalid entries in the correct format in the database.

VLOOKUP	FALSE
PostalCode	(Multiple Items)
Row Labels	**Count of Status**
Gone Out of Business	9
Inactive	10
Issued	648
Grand Total	**667**

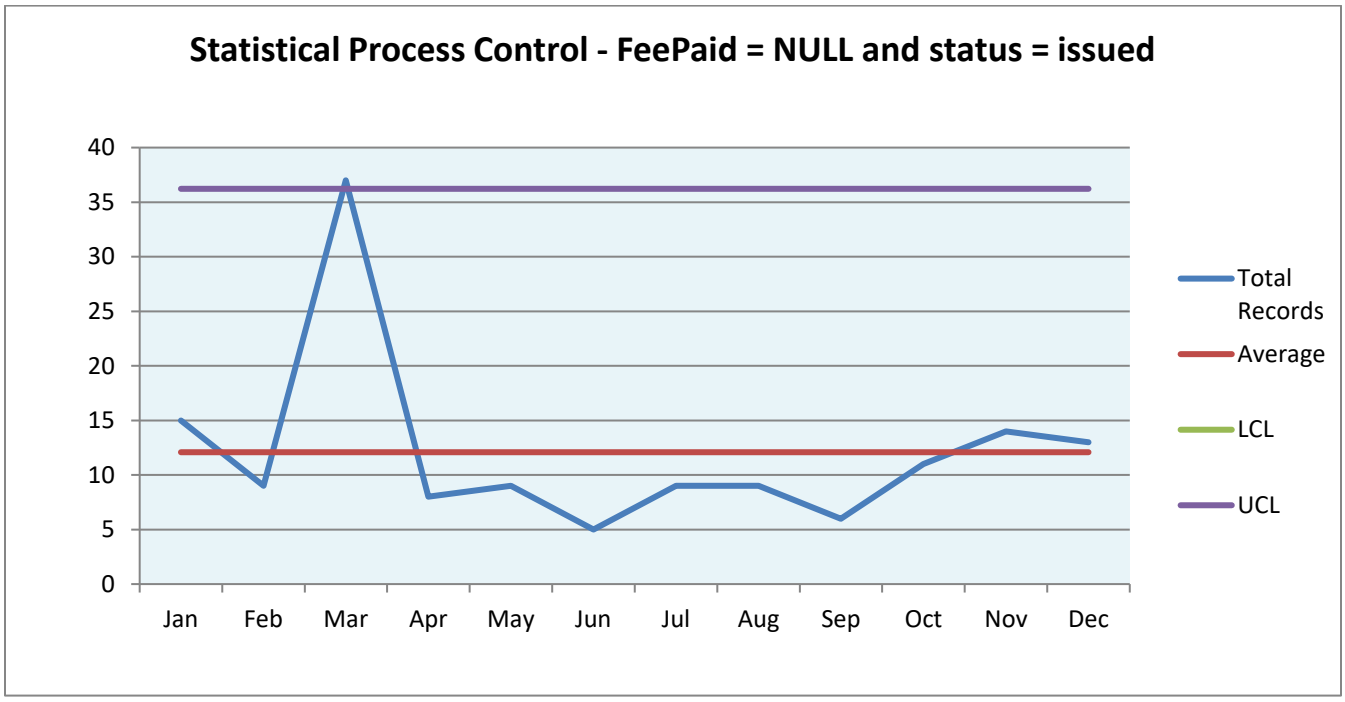
PostalCode: NULL values

The process breaches the upper bounds multiple times in November and December 2018 and is therefore out of control during those months. This process appears like the number of records being processed everyday as shown above.



FeePaid: NULL values

The process of recording the applicable fee is slightly out of control in the month of March 2018.

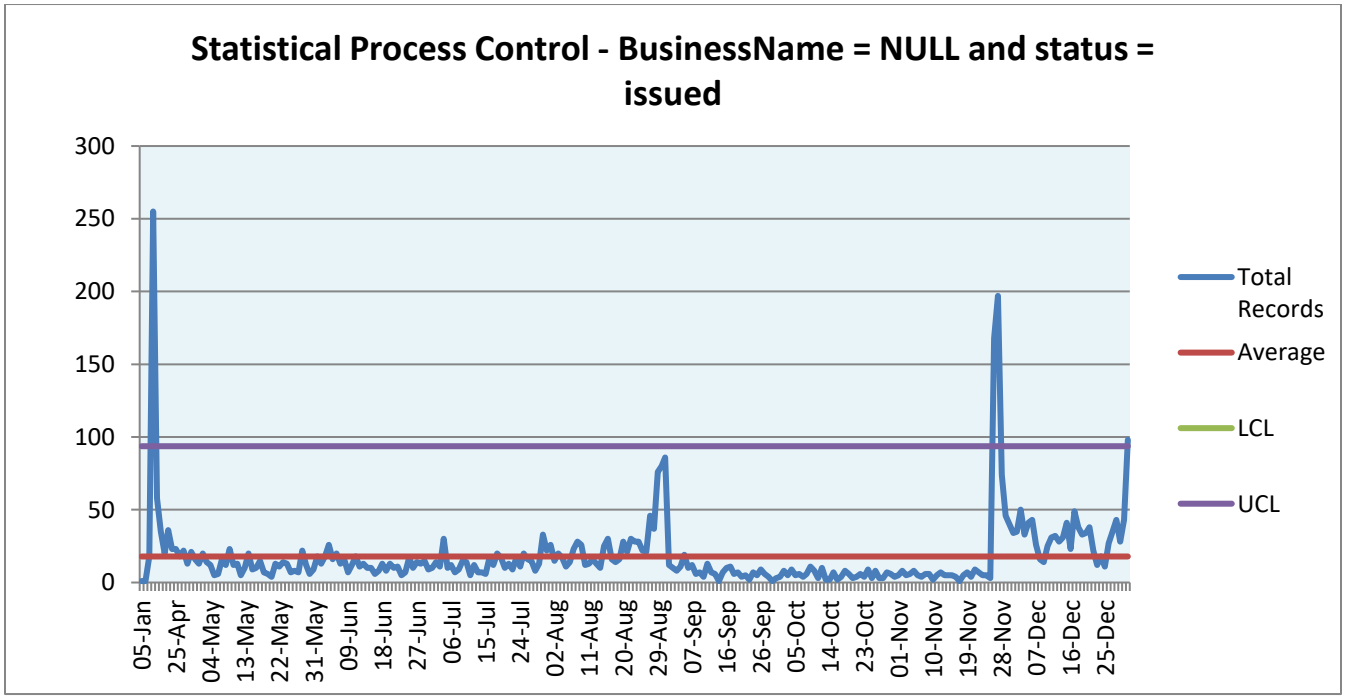


A further drill down based on the revision number indicates that most of the records where FeePaid = Null are from the first revision.

FeePaid	(blank)
Row Labels	**Count of Status**
0	**49**
Cancelled	5
Gone Out	11
Inactive	1
Issued	28
Pending	4
1	**114**
Gone Out	2
Inactive	3
Issued	109
2	**9**
Cancelled	1
Issued	8
Grand Total	**172**

BusinessName: NULL values

The process of recording business name as null when it is active is out of control for the months of January, November, and December 2018. As seen above, most of these records belong to the short-term business type.



SME Review of Further Research

LicenceRSN: Duplicates

The duplicate values are probably a one-off type of an error as the amount of data being inputted at the time of these errors was well within normal range. There are only two errors for the entire year however referential integrity must be enforced going forward.

PostalCode: Invalid records

The research conducted has helped define a valid record – one that is in a correct format and is a valid entry. A value in this column cannot be outside of the options available at a particular point in time. The scope of the problem is not large at the moment, however recording invalid values makes the database unreliable for analysis. The process seems to be out of control during periods of heavy data input which could be the cause of these errors.

PostalCode: NULL values

Further research into this matter shows that the process is out of control during the months of heavy data input. The resemblance between the two process control charts is high and is probably the reason for such errors.

FeePaid: NULL values

The control chart shows that the process is mostly within control except for the month of March. Further drill down based on licence version is helpful in determining the fact that most of these errors arise at the time of renewal or any other circumstance under which the licence version changes.

BusinessName: NULL values

The investigation into the matter has shown that these errors are concentrated in the short-term rental type of business. Since it highly likely that such businesses are registered under an individual, there is no registered company name to record. However, some indicator such as the name of the owner of the property should be recorded as business trade name. Also, the process is out of control during times of heavy data input which could also be the cause for such errors.

SME Suggests Some DQ Rules

LicenceRSN: Duplicates

Since the value in this field is to be used as a primary key, there should be a constraint on the field that does not allow duplicate values. There should be a prompt indicating that the user is about to commit an error if trying to input a value already in the database.

PostalCode: Invalid records

Going forward, it is proposed that only two formats (LDL DLD & DDDDD) be allowed if business is being conducted in Canada and US. As and when required, other formats may be added. The format allowed should be enforced based on the country selected and incorrect values should not be allowed in the database. Validity of the records can be strengthened by measuring the input against a current database of possible records. ‘Special values’ such as 000 000 & 00000 can be allowed as it is possible that a new postal code is added thereby making the existing database of possible options obsolete. These ‘special values’ can be further inputted into a statistical process control where status = issued. This can help identify the need for updating our database or other possible training issues.

PostalCode: NULL values

NULL value should not be allowed in this column when status = issued (active). If the business is active, we should have a current address including a valid postal code as defined above.

FeePaid: NULL values

NULL value should not be allowed in the column if the status=issued, however a value of zero can be allowed as it is possible that some business may qualify for a free licence/service.

BusinessName: NULL values

For status = active, records should not be allowed to have NULL values in both business name and business trade name fields. Some contexts to the data must be available in every record.