COMP3839

Masoomeh Sohrabi

Contents

Overview of Project

This project will assess a database of Vancouver Business License. The valid business licence is a requirement to run a business in Vancouver. The specific dataset for this exercise is for the year of 2017 of Vancouver Business License.

Improved data quality leads to better decision-making across an organization. The more high-quality data you have, the more confidence you can have in your decisions. Good data decreases risk and can result in consistent improvements in results.

Daniel Assad will be the SME for the project because he has been managing the Vancouver Business Licencing system for 10 years and understands the data and processes thoroughly. The main goal of this project is to look at all data in the database to find the issues in the columns and fix them to improve the report and have a better analyzing of data.

In the initial assessments, the quality issues in 5 selected columns are determined and the SME will review the initial assessments and then will suggest some data rules to fix or prevent the problems. In the SME further research and the statistical of the date will be reviewed using the Statically Process Chart to analysis the date of the area that are not under control.

Here is the link of the data set:

https://opendata.vancouver.ca/explore/dataset/businesslicences/export/?disjunctive.status&disjunctive.businesssubtype&refine.issueddate=2017

The columns in the dataset are described as in the table below:

Column Name	Data type	Description
FOLDERYEAR	String	First 2characters of the business license number, representing the expired date
LisenceRSN	String	It is a unique number for each business that generate by the system
LicenceNumber	String	It is a 9-character that 2-digit indicating the expired year followed by a hyphen and the 6-digit number that generated by the system.
LicenseRevisionNumber	String	This is a 2-digit number?
BusinessName	String	This indicates the Name of the business
Businesstradename	String	This is the name under which business is usually conducted
Status	String	This shows the status of the business license.
BusinessType	String	This describes the nature of the business and what the business is about.
IssedDate	String	The date when the business license issued.
ExpieredDate	String	The date the business license expired.
BusinessType	String	It is the description of the business activities.
BusinessSubType	String	It is the subcategory of the main business type.
Unit	String	This is the building identifier in the mailing address.
UnitType	String	A description of a location
House	String	The number assigned to an address where the business is located.
Street	String	The name of the street where the business is located
City	String	Name of the municipality where the business is located
Province	String	Name of the province or state where the business is located
Country	String	A 2-character that indicates the country where the business is located
PostalCode	String	The postal code of place where the business is located
LocalArea	String	Manual selection from data custodian in source system.
NumberofEmployees	String	Number of staff employed with the business
FeePaid	String	The total amount of license fee paid in Canadian Dollar
ExtractDate	String	Date when data was extracted from source data system.
Geom	String	Special representation of feature

Initial Assessment

1. LicenceRSN : Duplicates

In the LicenseRSN column has 15 duplicated records as shown in the basic tab also in the frequency tab you can see the IDs are duplicated. The LisenceRSN is to be a Unique number because is a primary key for this table.

Type	Count	%
Null	0	0.00%
Non-null	51,419	100.00%
Duplicate	15	0.03%
Distinct	51,404	99.97%
Non-uni...	15	0.03%
Unique	51,389	99.94%

Frequency Analysis

Range: none

100 most common values:

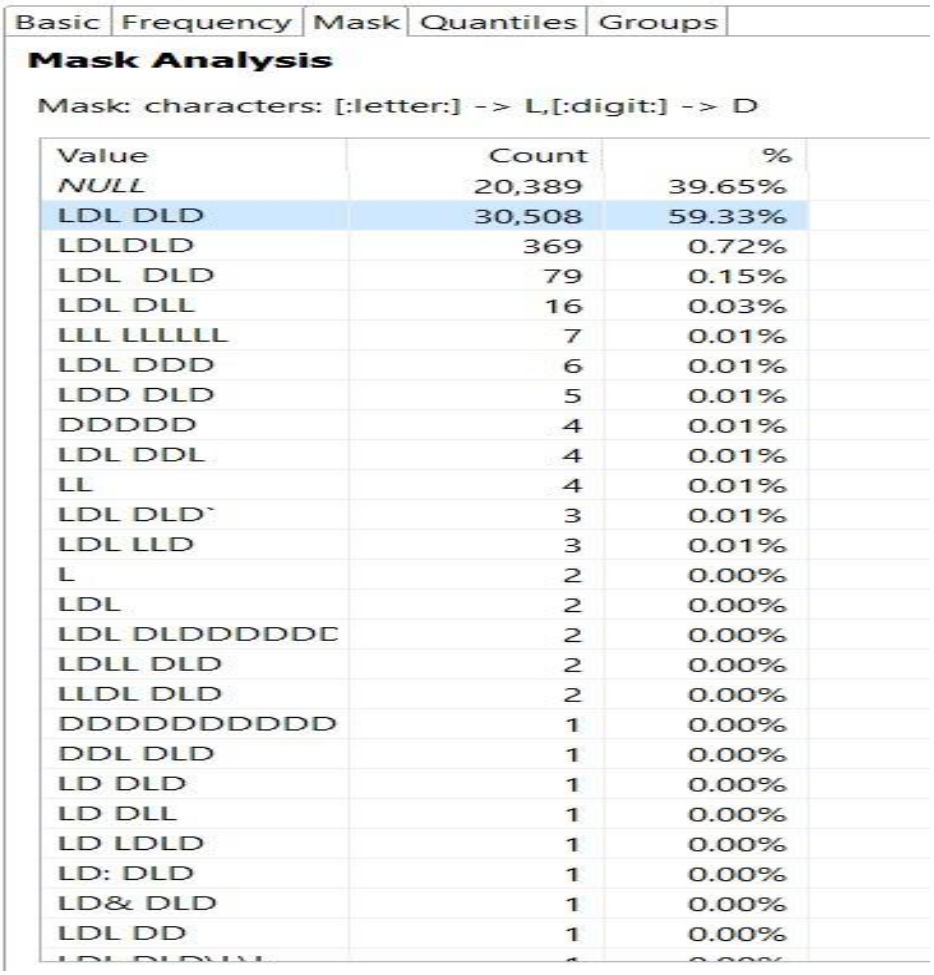
Value	Count	%
2921611	2	0.00%
3012422	2	0.00%
3012423	2	0.00%
3012424	2	0.00%
3012425	2	0.00%
3012426	2	0.00%
3012427	2	0.00%
3012428	2	0.00%
3012429	2	0.00%
3012430	2	0.00%
3012431	2	0.00%
3012432	2	0.00%
3012433	2	0.00%
3012434	2	0.00%
3012557	2	0.00%
2130659	1	0.00%
2570128	1	0.00%
2580429	1	0.00%

By looking at below table, you can find a sample of 15 duplicate records. We can see that most columns of duplicate records have same value; however, those have different addresses. Also, you can see except of first one the others are coming in the row and have same Issued date(day).

	BusinessName	Status	IssuedDate	Street	City	Province	Country	PostalCode
2921611	Sang Eun Lee (Sang Lee)	Issued	2017-05-18T13:41:30-07:00	W HASTINGS ST	Vancouver	BC	CA	V6E 4R5
2921611	Sang Eun Lee (Sang Lee)	Issued	2017-05-18T13:41:30-07:00	Acadia Road	Vancouver	BC	CA	V6T 1R3
3012422	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:47-08:00	W 41st Av	Vancouver	BC	CA	V6M 2A4
3012422	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:47-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012423	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:47-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012423	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:47-08:00	Terminal Av	Vancouver	BC	CA	V6A 4G2
3012424	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:48-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012424	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:48-08:00	W 4th Av	Vancouver	BC	CA	V6K 1N9
3012425	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:48-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012425	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:48-08:00	W Broadway	Vancouver	BC	CA	V6R 2B1
3012426	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:49-08:00	Commercial Dr	Vancouver	BC	CA	V5L 3Y3
3012426	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:49-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012427	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:49-08:00	E Hastings St	Vancouver	BC	CA	V5K 1Z3
3012427	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:49-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012428	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:50-08:00	Kingsway	Vancouver	BC	CA	V5R 5K6
3012428	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:50-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012429	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:50-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012429	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:50-08:00	Main St	Vancouver	BC	CA	V5V 3P8
3012430	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:50-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012430	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:50-08:00	W 10th Av	Vancouver	BC	CA	V5Z 1K9
3012431	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:51-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012431	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:51-08:00	Dunbar St	Vancouver	BC	CA	V6A 4G2
3012432	Vancouver City Savings Credit Union	Issued	2017-12-05T13:58:46-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012432	Vancouver City Savings Credit Union	Issued	2017-12-05T13:58:46-08:00	Fraser St	Vancouver	BC	CA	V5W 3A4
3012433	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:51-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012433	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:51-08:00	W Pender St	Vancouver	BC	CA	V6C 1J8
3012434	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:52-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012434	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:52-08:00	Victoria Dr	Vancouver	BC	CA	V5P3W1
3012557	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:52-08:00	TERMINAL AV	Vancouver	BC	CA	V6A 4G2
3012557	Vancouver City Savings Credit Union	Issued	2017-12-05T14:09:52-08:00	W 10th Av	Vancouver	BC	CA	V6R 4N2

2. Postal code : Invalid Format and Null

The standard format for Canadian postal code is LDL DLD and standard format for American postal code is either 99999 or 999999999. As shown in the table below, there are many different format values rather that Canadian and American valid standards.



Basic | Frequency | Mask | Quantiles | Groups

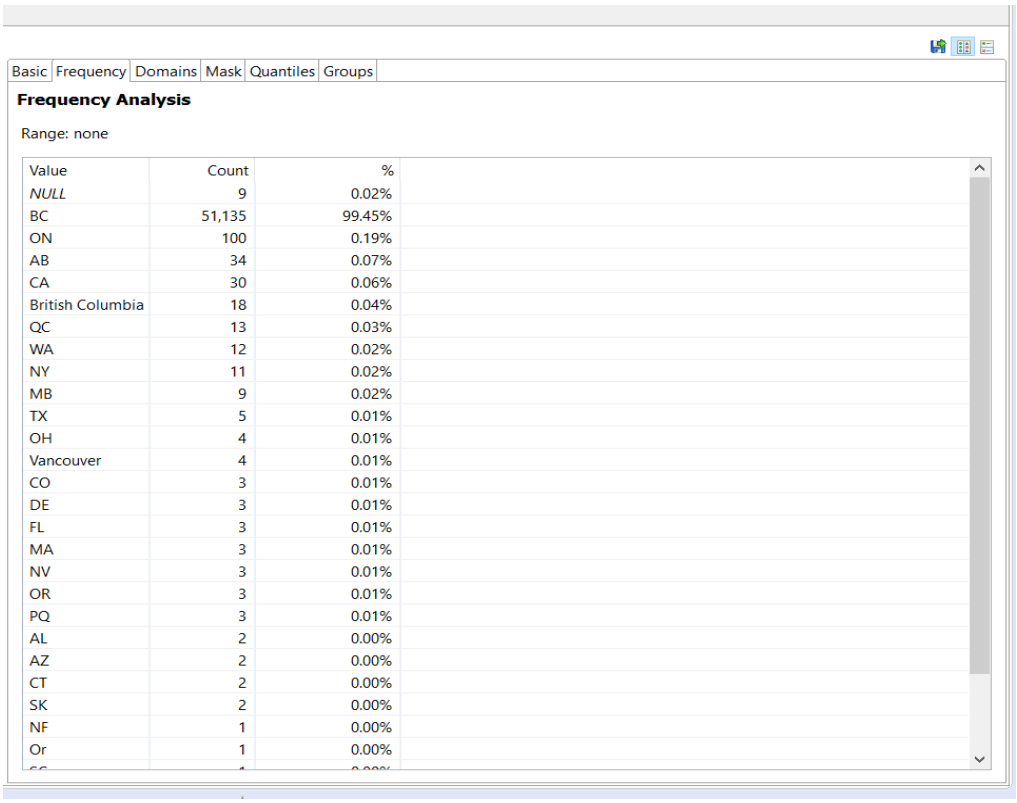
Mask Analysis

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

Value	Count	%
NULL	20,389	39.65%
LDL DLD	30,508	59.33%
LDLDLD	369	0.72%
LDL DLD	79	0.15%
LDL DLL	16	0.03%
LLL LLLLLL	7	0.01%
LDL DDD	6	0.01%
LDD DLD	5	0.01%
DDDDD	4	0.01%
LDL DDL	4	0.01%
LL	4	0.01%
LDL DLD`	3	0.01%
LDL LLD	3	0.01%
L	2	0.00%
LDL	2	0.00%
LDL DLDDDDDD	2	0.00%
LDLL DLD	2	0.00%
LLDL DLD	2	0.00%
DDDDDDDDDD	1	0.00%
DDL DLD	1	0.00%
LD 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%

3. Province : Two letter standard not followed

The basic can determine there are different value format for province. There are some entries with the full name of province and some with 2 letters. As it is shown in the frequency tab you can see the Vancouver and British Colombia are entered into province column.



Basic | Frequency | Domains | Mask | Quantiles | Groups

Frequency Analysis

Range: none

Value	Count	%
NULL	9	0.02%
BC	51,135	99.45%
ON	100	0.19%
AB	34	0.07%
CA	30	0.06%
British Columbia	18	0.04%
QC	13	0.03%
WA	12	0.02%
NY	11	0.02%
MB	9	0.02%
TX	5	0.01%
OH	4	0.01%
Vancouver	4	0.01%
CO	3	0.01%
DE	3	0.01%
FL	3	0.01%
MA	3	0.01%
NV	3	0.01%
OR	3	0.01%
PQ	3	0.01%
AL	2	0.00%
AZ	2	0.00%
CT	2	0.00%
SK	2	0.00%
NF	1	0.00%
Or	1	0.00%

4. Feepaid: Null value

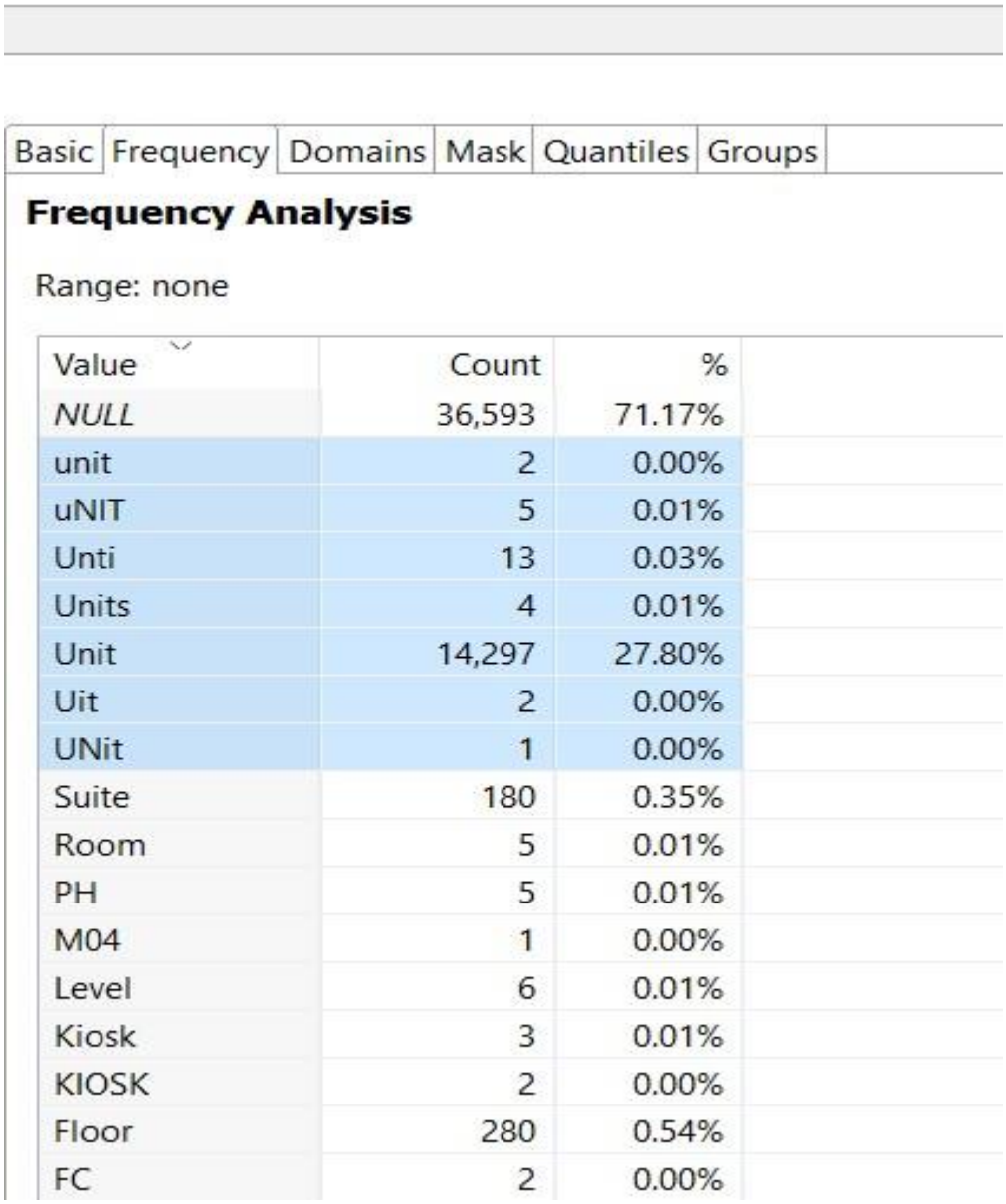
The basic tab in FeePaid column indicates 228 records of fee paid are null. If the null value for feepaid column is for the licence number that have the issued status, so it means there many businesses that have not paid their fee even though they are issued. Count of Null FeePaid records for each Status are shown in below table. Issued Status has maximum count with 196 records out of 228.

Type	Count	%
Null	228	0.44%
Non-null	51,191	99.56%
Duplicate	49,624	96.51%
Distinct	1,567	3.05%
Non-uni...	821	1.60%
Unique	746	1.45%

Status	Count of FeePaid = Zero/Null
Cancelled	7
Gone Out of Business	13
Inactive	8
Issued	196
Pending	4

5. UnitType: Variation on Unit Value

In the UnitType column, the variation on “Unit” is high. There is some value that are not meaningful. As you can see in the frequency table in the below, there some different spelling of “unit”. Since the only valid value is “Unit” and the rest of them such as uNit, Unti, Untis, Uit ,… are not acceptable in the address so this issue is better to be fixed.



Basic | Frequency | Domains | Mask | Quantiles | Groups

Frequency Analysis

Range: none

Value	Count	%
NULL	36,593	71.17%
unit	2	0.00%
uNIT	5	0.01%
Unti	13	0.03%
Units	4	0.01%
Unit	14,297	27.80%
Uit	2	0.00%
UNit	1	0.00%
Suite	180	0.35%
Room	5	0.01%
PH	5	0.01%
M04	1	0.00%
Level	6	0.01%
Kiosk	3	0.01%
KIOSK	2	0.00%
Floor	280	0.54%
FC	2	0.00%

SME Review of Initial Assessment

1. LicenceRSN : Duplicates _High priority

As LicenseRSN is defined as a unique identifier for each record so its value must be unique, and all records must have different LicenseRSN values. So, this issue has a high priority and needs to fix immediately.

2. Postal code: invalid format _High priority.

The standard format for the Canadian postal code should be LDL DLD and any format that matches this standard is valid. Only %59 of the postal code is in the right format. As the wrong postal will result return the important documents such as invoices that will impact the financial situation. So, the priority of fixing this issue is high

3. Province : Two letter standard not followed _Low priority

In this database, the province is indicating by two standard letters. There are some entries that not following the 2-standard level. (British Colombia, Vancouver and Or). The correct format value for this data supposed to be LL but as it is shown in the Mask tab, there are other value format.

As %99.94 of the data is in the correct format and this it does not impact the issuing the business license so the priority for fixing this issue is low.

4. FeePaid: Null- High Priority

The greatest number of null feePaid belongs to issued Status which are the businesses that they already have their Licence and the fee supposed to be paid. Since this issue could cause a financial problem; therefore, we need to discover the cause of this issue. So, the priority for fixing this issue is High.

5. UnitType: Variation on Unit- Low Priority

It is better the unit type of the business location be determined. In this data base some unit type that entered are not acceptable. However, the unit type is not a mandatory field in address line and is not critical for the accurate address, so the priority of fixing this data is low.

Further Research for the SME

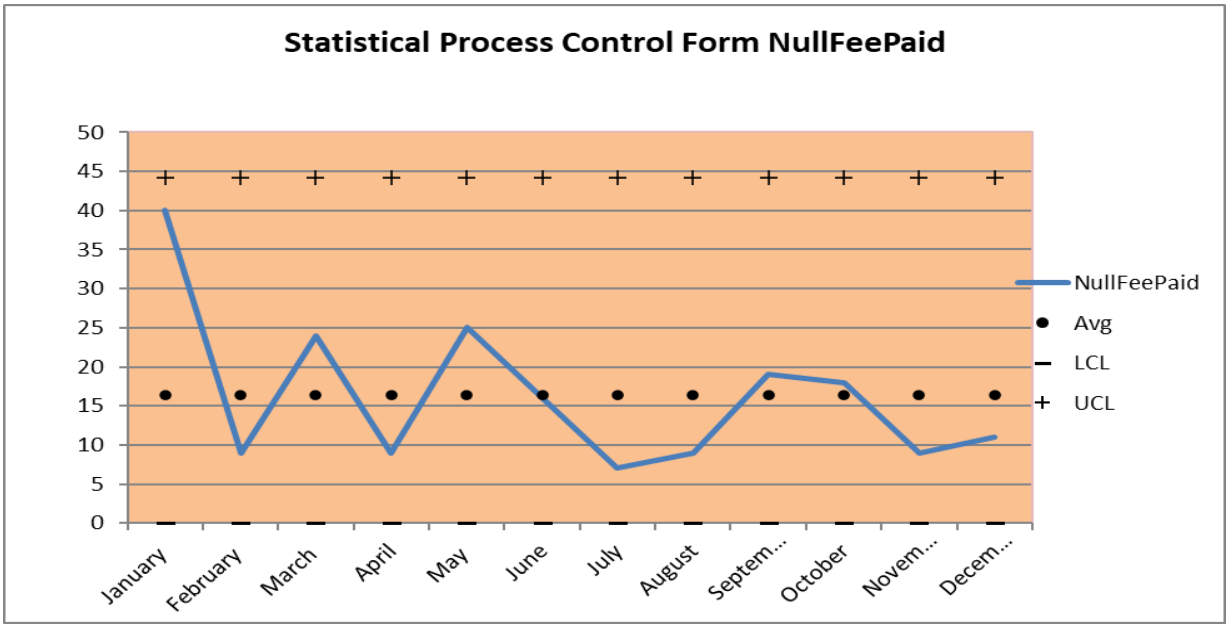
For the further research for the SME, the null issue on FeePaid column will be reviewed. The statistical Process chart will be using to determine if the data quality process is in a state of control.

For this purpose, we need a count by month of Issued Date where FeePaid are null, and the Status is issued. we have 196 records.

To have better understanding of null feepaid, this table provides statistics of null number of this field between January and December for 2017.

type	Month	Count	AVG	LCL	UCL
NullFeePaid	January	40	16	0.00	44
NullFeePaid	February	9	16	0.00	44
NullFeePaid	March	24	16	0.00	44
NullFeePaid	April	9	16	0.00	44
NullFeePaid	May	25	16	0.00	44
NullFeePaid	June	16	16	0.00	44
NullFeePaid	July	7	16	0.00	44
NullFeePaid	August	9	16	0.00	44
NullFeePaid	September	19	16	0.00	44
NullFeePaid	October	18	16	0.00	44
NullFeePaid	November	9	16	0.00	44
NullFeePaid	December	11	16	0.00	44

Form Type	Average	StDev
NullFeePaid	16.33	9.29



SME Review of Further Research

With the review of the Statistical Process Chart, null value is the issue that was happening in the whole year of 2017 between January till December. It means that nothing is out of control and our system working properly.

SME Suggests Some DQ Rules

LicenceRSN

LicenceRSN is a primary key and so it is uniqueness should be enforced in the application or database. We can put the uniqueness constraint for the column, which would not allow for duplicate values We need to talk to database developer for set up more detailed system check about duplicates.

PostalCode:

PostalCode must match the format of the Country for the record if the Status=Issued. So depended on the country it is better to force a default format for postal code in application layer, so no more invalid postal code will store in related database.

Province:

Province must be indicated as a two-capital letter. We can Format and limit the character allowed to be entered to two. It would be an option to use the drop-down feature instead of entering the province manually.

FeePaid:

For preventing the Null FeePaid, it is better to have default value like -1 for Pending Status or set a meaningful default value for other Statuses as well if it is required; Therefore, for providing more accuracy in financial reports, the Null values that already have entered in database should replace with proper default value.

UnitType:

We must introduce the “Unit” value in the Unittype column in the database and force to use the exist value from a drop- down feature.