

Greater Vancouver Business License

Banafsheh Hassani

2019-04-06

LinkedIn: <https://www.linkedin.com/in/banafsheh-hassani-7b063a129/>

Github: <https://github.com/BanafshehHassani>

Contents

Overview of Project	2
Initial Assessment	4
LicenseRSN: Duplicated Data in Primary Key.....	4
UnitType: Un-Normal Data	6
City: Un-Normal Data	8
Country: Abbreviations.....	10
PostalCode: Invalid Postal Code.....	11
SME Review of Initial Assessment	15
FeePaid : Lack of the information	15
Further Research for the SME	15
SME Review of Further Research	16
SME Suggests Some DQ Rules	16

Overview of Project

Valid license is required in order to operate a business in Vancouver, declared By-Law No. 4450, available in <https://data.vancouver.ca/datacatalogue/businessLicence.htm>. The business licence can be obtained from the City's Licence Office and would be valid for the remainder of the calendar year, unless stated otherwise. The investigated dataset for this project has been collected

from the above-mentioned online source, including all business licence records issued within the Greater Vancouver since 1997.

The studied dataset in this project provide information on the businesses licence number, name, type, headquarter location, number of hired employee, current status, operation history, business initiation, shut down time and the fee paid (if available). Initial screen of the dataset revealed some anomalies (e.g. blank cells, duplicates, abbreviations etc.), which has been extensively analyzed and discussed in the next sections of this report. The overarching objective of this project is to identify the potential anomalies in order to increase the data quality for any required future processing. Microsoft Excel 2019 was employed in order to perform the data quality assessment (DQA) including an initial assessment SME review, further SME research and review, and the SME suggestions via DQ rules.

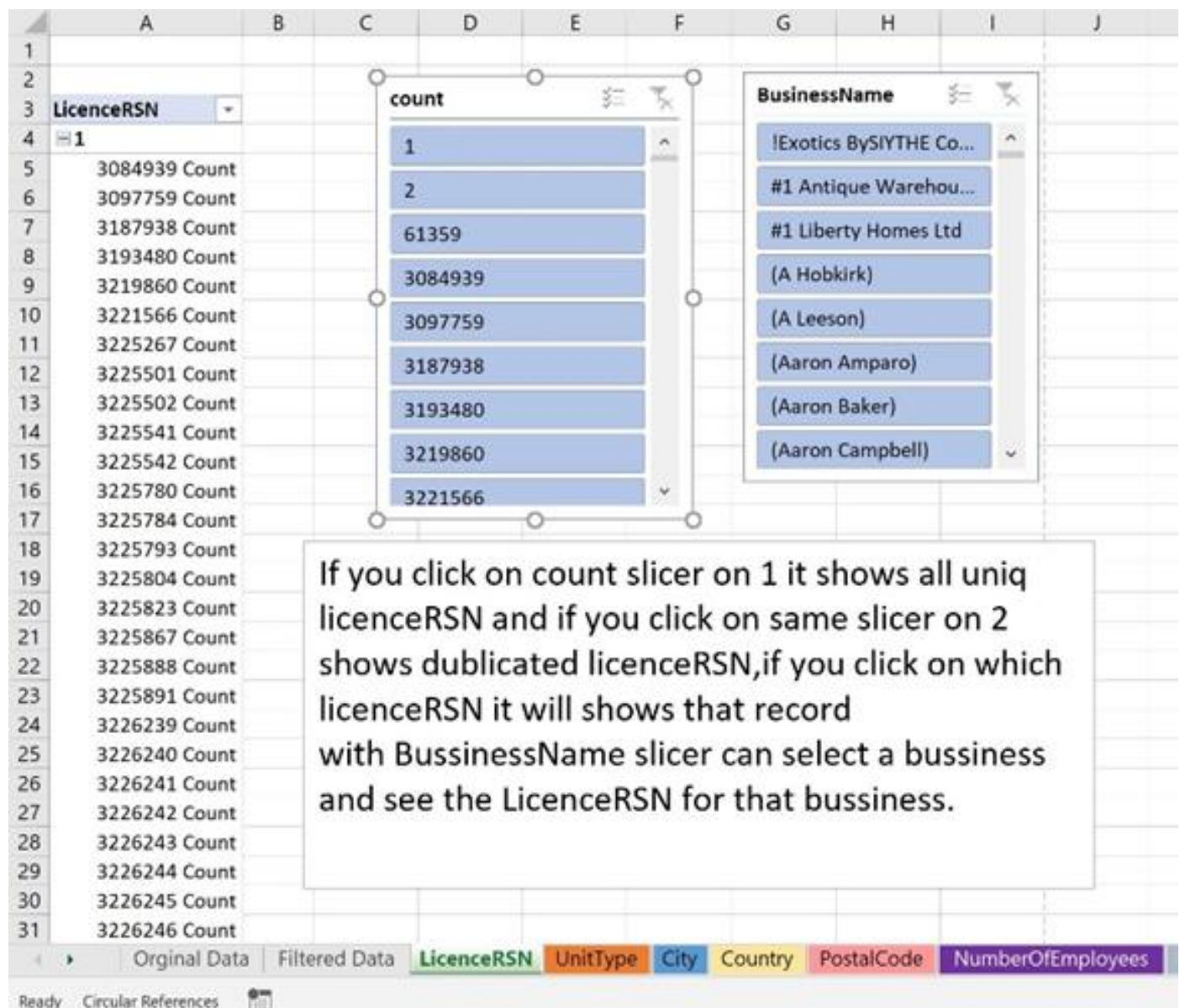
Initial Assessment

LicenseRSN: Duplicated Data in Primary Key

Initial evaluation of the LicenseRSN column revealed that there are a number of duplicated records within the studied dataset of 122716 rows. In addition, there are some blank cells presented in the data. More in detail information about the duplicated and blank data can be found via the attached Excel file, under the LicenseRSN sheet, using slicer described in below figure.

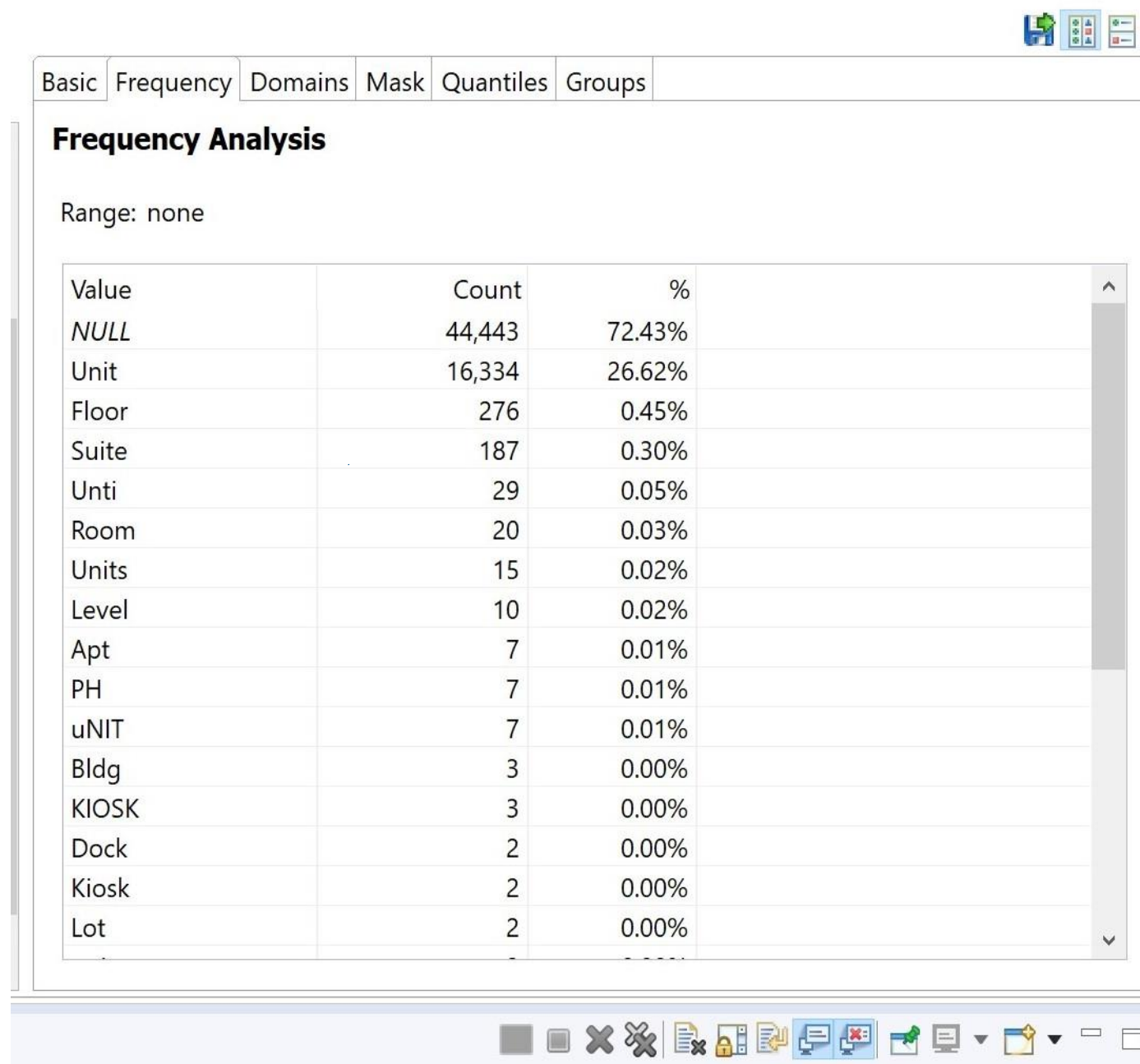
Duplicated LicenceRSN

	A	B	C	D	E	F	G	H	I	
1	Count	LicenceRSN	LicenceNur		BusinessName	BusinessTradeName	Status	IssuedDate	ExpiredDate	But
239	2	3346585	19-208281	00	SAGE Dining Services Canada Ltd		Pending			Pe
4850	2	3235769	19-110257	00	1003259 BC Ltd	Cho Construction	Pending			Co
4851	2	3235769	19-110257	00	1003259 BC Ltd	Cho's Construction	Pending			Co
4852	2	3235770	19-110258	00	1003259 BC Ltd	Cho Construction	Pending			Co
4853	2	3235770	19-110258	00	1003259 BC Ltd	Cho's Construction	Pending			Co
7084	2	3250639	19-125069	00	Bring to Balance Inc	Bring to Balance's Art For Change Centre	Pending			He
7085	2	3250639	19-125069	00	Bring to Balance Inc	Bring to Balance's Art For Change Centre	Pending			He
3882	2	3254886	19-129275	00	Sun 8 Holdings Inc		Pending			Ol
3883	2	3254886	19-129275	00	Sun 8 Holdings Inc		Pending			Ol
3077	2	3271836	19-146077	00	Snowdeer Jewelry Inc		Issued	2018-12-18 14:05:54	2019-12-31 00:00:00	Ri
3078	2	3271836	19-146077	00	Snowdeer Jewelry Inc		Issued	2018-12-18 14:05:54	2019-12-31 00:00:00	Ri
3727	2	3336640	19-198393	00	Trong H Nguyen & Mien T Nguyen & Thi N Bui	Mien T Nguyen (Mien Nguyen)	Issued	2019-02-21 09:30:07	2019-12-31 00:00:00	Ol
3728	2	3336640	19-198393	00	Trong H Nguyen & Mien T Nguyen & Thi N Bui	Trong H Nguyen (Trong Nguyen)	Issued	2019-02-21 09:30:07	2019-12-31 00:00:00	Ol
1301	2	3346585	19-208281	00	SAGE Dining Services Canada Ltd		Pending			Pe



UnitType: Un-Normal Data

The initial assessment on the UnitType column showed that some of the records are filled with number while a number of cells in this column are completely blank. More information can be found in the attached Excel file (UnitType sheet) with a general overview available in below screenshots.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
3	UnitType		BusinessName			UnitType													
4	0					0													
5	1		IExotics BySIYTHE C...			1													
6	2		#1 Antique Wareho...			2													
7	16916		#1 Liberty Homes Ltd			16916													
8	550		(A Hobkirk)			550													
9	Apt		(A Leeson)			Apt													
10	Bay		(Aaron Amparo)			Bay													
11	Bldg		(Aaron Baker)			Bldg													
12	Dock		(Aaron Campbell)																
13	Dock																		
14	F12																		
15	FC																		
16	Floor																		
17	INCL																		
18	Kiosk																		
19	Level																		
20	Lot																		
21	PH																		
22	Room																		
23	Suite																		
24	Uit																		
25	Unit																		
26	Unit																		
27	Unit'																		
28	Units																		
29	Unti																		
30	Untis																		
31	(blank)																		
32	Grand Total																		
33																			

in the UnitType, can see some data anomalies, one of the most important of them is some number including 0,1,2 and the lack of Unitype for some rows is obvious. However, these two sliders show the anomalies just pointed.

City: Un-Normal Data

First assessment on the City column demonstrated that there are a few cells filled with numbers. In addition, there are a number of blank cells available in this column. Also it has many example of two or more spellings of a city name. Below two figures illustrate an overlook on these anomalies while more information can be reached via the attached Excel file (City sheet).

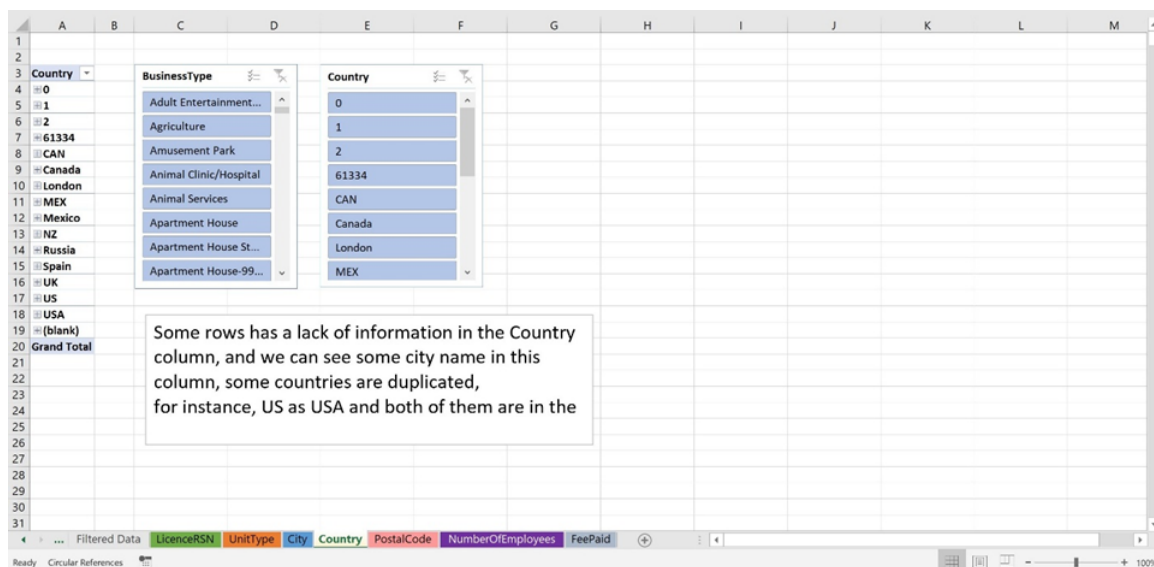
<u>City Name</u>	<u>Number of business</u>
0	61355
1	1
2	0
61331	1
Vancouver	55125
100 Mile House	1
301	1
Abbotsford	197
Abbotsford	4
Burnaby	924
Burnbaby	1

As you can see City column has some wrong dataes such as number for the city name and some rows has a blank city name, which doesn't make sense. if you click on city slicer and choose what city you want you can see all the business are exited in that city, which doesn't make sense for city 1 or blank.

Country: Abbreviations

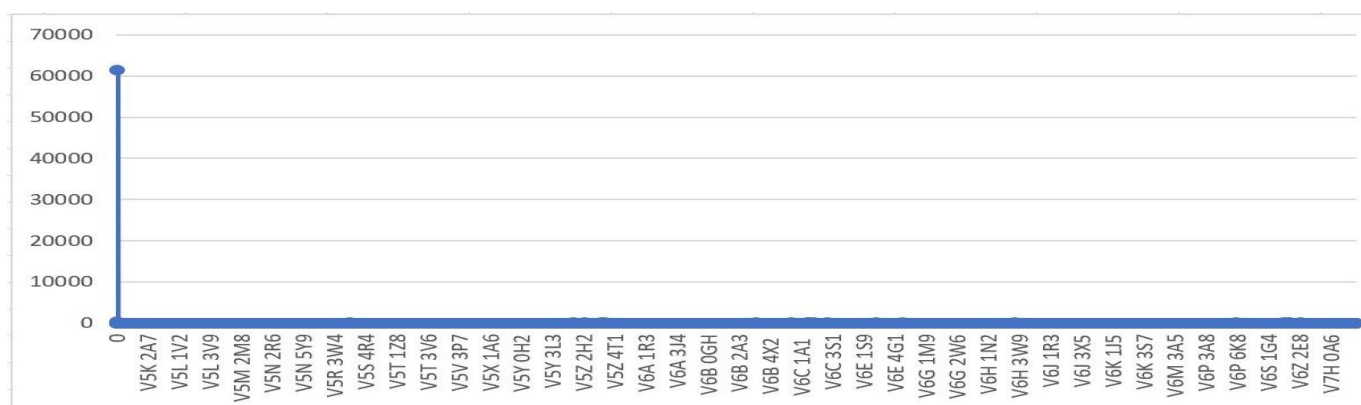
Initial analysis of the Country column demonstrated that there are a number of cases where a country is called by both full name and abbreviation in the data records. For instance, Canada and USA are also referred as Can and Us in this column, respectively.

<u>Country</u>	<u>Number of Business</u>
0	61355
1	1
2	0
61334	1
CAN	57824
Canada	3443
London	1
MEX	1
Mexico	1
NZ	1
Russia	1
Spain	1
UK	1
US	1
USA	59



PostalCode: Invalid Postal Code

The first screening of the data shows that there are a few incorrect postal code format in which include number and characters (e.g. single character/number of 1 and a). The latter might make some troubles in the next analyzing steps together with potential for security issues.



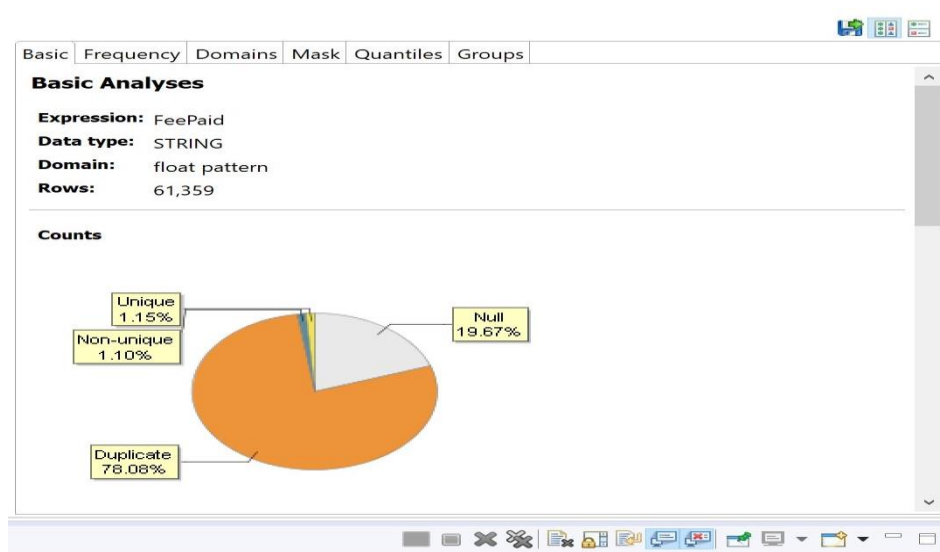
Mask Analysis

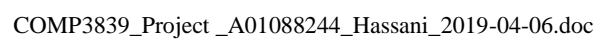
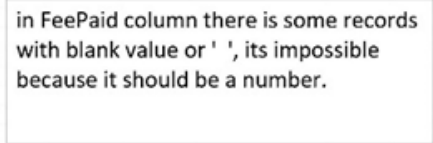
Mask: characters: [:letter:] -> L[:digit:] -> D

Value	Count	%
<i>NULL</i>	28,125	45.84%
LDL DLD	32,772	53.41%
LDLDLD	357	0.58%
LDL DLD	20	0.03%
LDL DLL	16	0.03%
LLL LLLLLL	12	0.02%
LDD DLD	8	0.01%
LDL DDD	6	0.01%
LDL LLD	5	0.01%
DDDDD	4	0.01%
LDL DDL	4	0.01%
LD DLD	3	0.00%
LDL	3	0.00%
L	2	0.00%
LD LDLD	2	0.00%
LDL)LD	2	0.00%

FeePaid : NULL Values

Initial analysis of the FeePaid column demonstrated that there exist 19.67% cells is Null. This might make some categorizing and analyzing issue in the next step of data analyzing.





SME Review of Initial Assessment

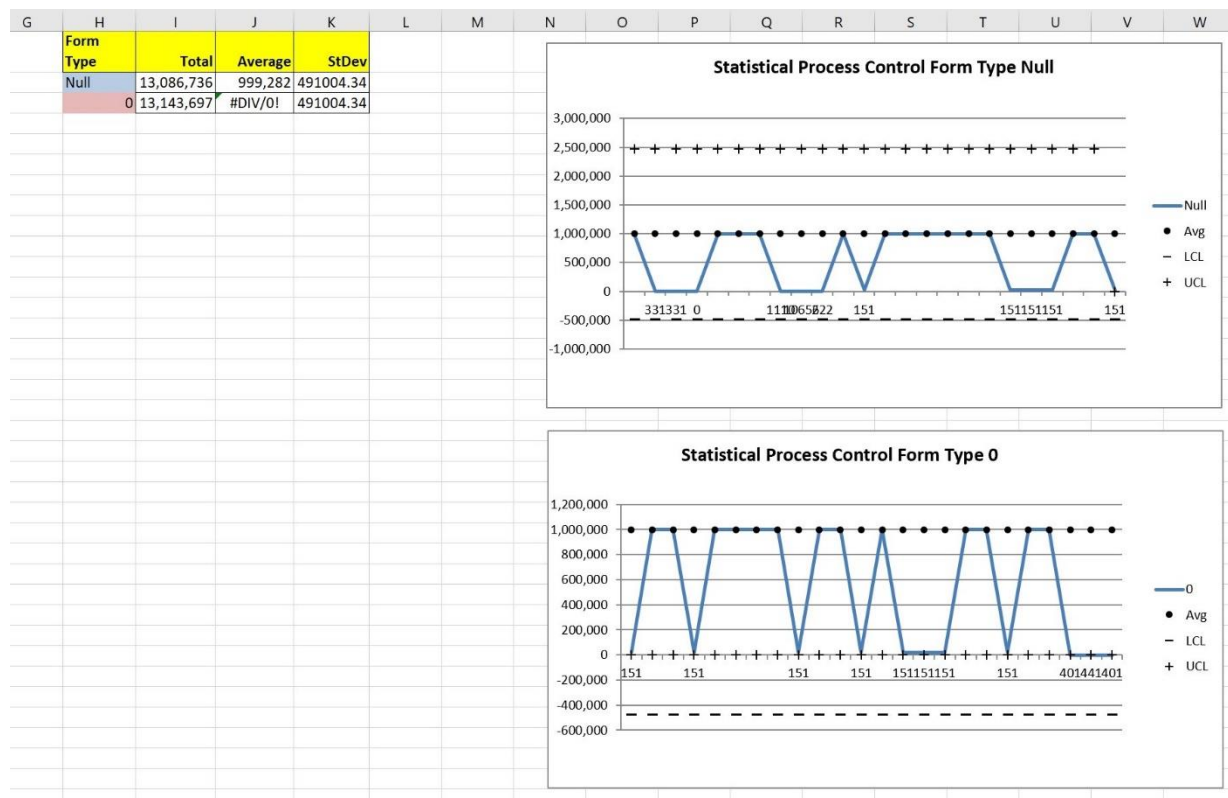
FeePaid : Lack of the information

Initial screening of the FeePaid column demonstrated that there are some Null value cells in this column. The latter can possibly increase the complexity of categorizing and analyzing. Therefore, it might be considered as on of the **most important priority subject**.

[illegible]

Further Research for the SME

In the FeePaid column has identified some anomalies (e.g. Null and '0') together with proposing an efficient solution pathway.



SME Review of Further Research

Further research was conducted on the potential anomalies in the FeePaid columns of the investigated dataset. A number of Zero and null data were found in the FeePaid column. This issue might be the most important issue in the next steps of data categorization and analysis owing to making several difficulties for registered licenses. In addition, massive data recollection would be required for FeePaid columns where majority of cells are filled with a value of null or zero.

SME Suggests Some DQ Rules

The most important challenge is on the existence of **duplicated** cells in the LicenseRSN column as the **primary key**. The latter has to be **discussed with IT department** in order to find the most efficient way for making the primary key unique.

UnitType and City columns have some **un-normal data** which should be resolved by **recollecting data** information from majority of the registered businesses. There are also a number of cities listed in dataset but located outside of the greater Vancouver. These rows should also be **removed** in order to **improve the data quality**.

With regards to the Country column, there are some items which have been referred as **both full name and abbreviations**. There is also a number of country cells filled with the **city name**. In addition, Postalcode, Numberofemployee and FeePaid columns include some data which seems to be **not in the correct format** to be used in their column. The latter might be **fixed via reformatting and in some cases, recollecting the missing data**.