Greater Vancouver Business License

Banafsheh Hassani

2019-04-06

Contents

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.

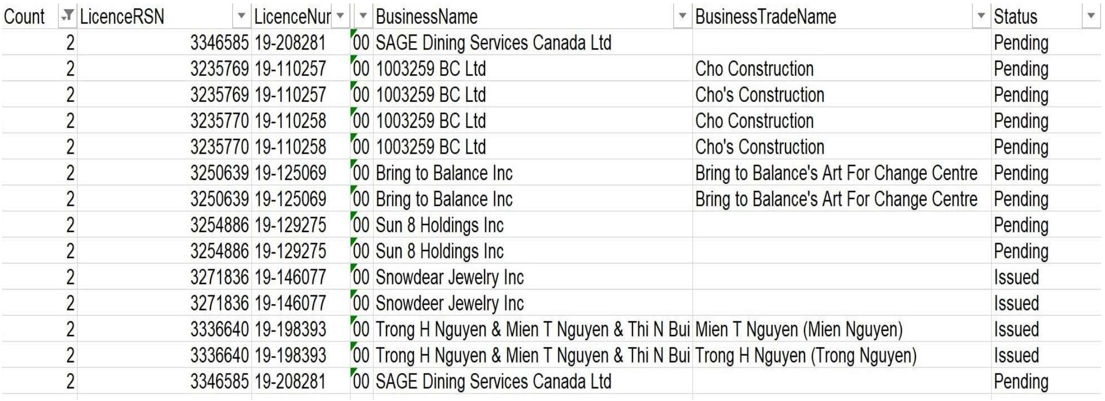
The studied dataset in this project provide information on the businesses licence number, name, type, headquarter location, number of hired employee, current status, operation history, business initiation, shut down time and the fee paid (if available). Initial screen of the dataset revealed some anomalies (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

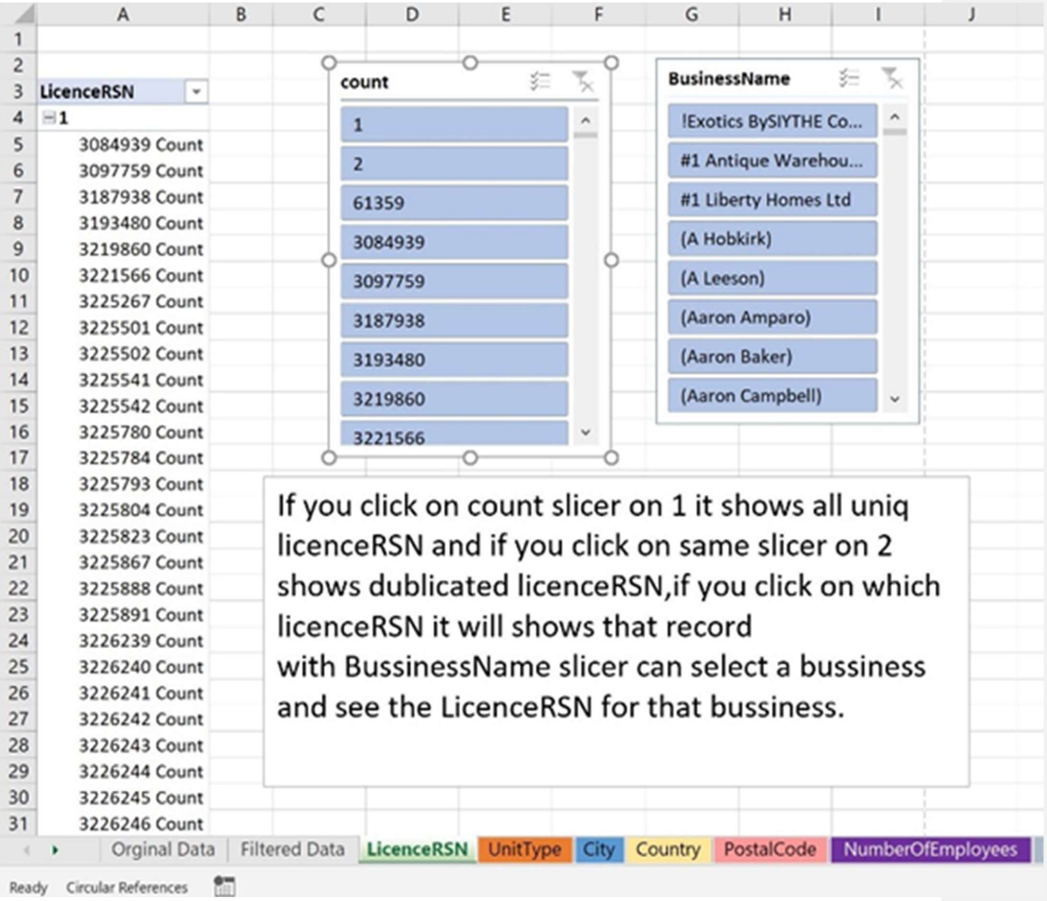
LicenseRSN: Duplicated Data in Primary Key

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



Count	LicenceRSN	LicenceNur		BusinessName	BusinessTradeName	Status
2	3346585	19-208281	00	SAGE Dining Services Canada Ltd		Pending
2	3235769	19-110257	00	1003259 BC Ltd	Cho Construction	Pending
2	3235769	19-110257	00	1003259 BC Ltd	Cho's Construction	Pending
2	3235770	19-110258	00	1003259 BC Ltd	Cho Construction	Pending
2	3235770	19-110258	00	1003259 BC Ltd	Cho's Construction	Pending
2	3250639	19-125069	00	Bring to Balance Inc	Bring to Balance's Art For Change Centre	Pending
2	3250639	19-125069	00	Bring to Balance Inc	Bring to Balance's Art For Change Centre	Pending
2	3254886	19-129275	00	Sun 8 Holdings Inc		Pending
2	3254886	19-129275	00	Sun 8 Holdings Inc		Pending
2	3271836	19-146077	00	Snowdear Jewelry Inc		Issued
2	3271836	19-146077	00	Snowdeer Jewelry Inc		Issued
2	3336640	19-198393	00	Trong H Nguyen & Mien T Nguyen & Thi N Bui	Mien T Nguyen (Mien Nguyen)	Issued
2	3336640	19-198393	00	Trong H Nguyen & Mien T Nguyen & Thi N Bui	Trong H Nguyen (Trong Nguyen)	Issued
2	3346585	19-208281	00	SAGE Dining Services Canada Ltd		Pending

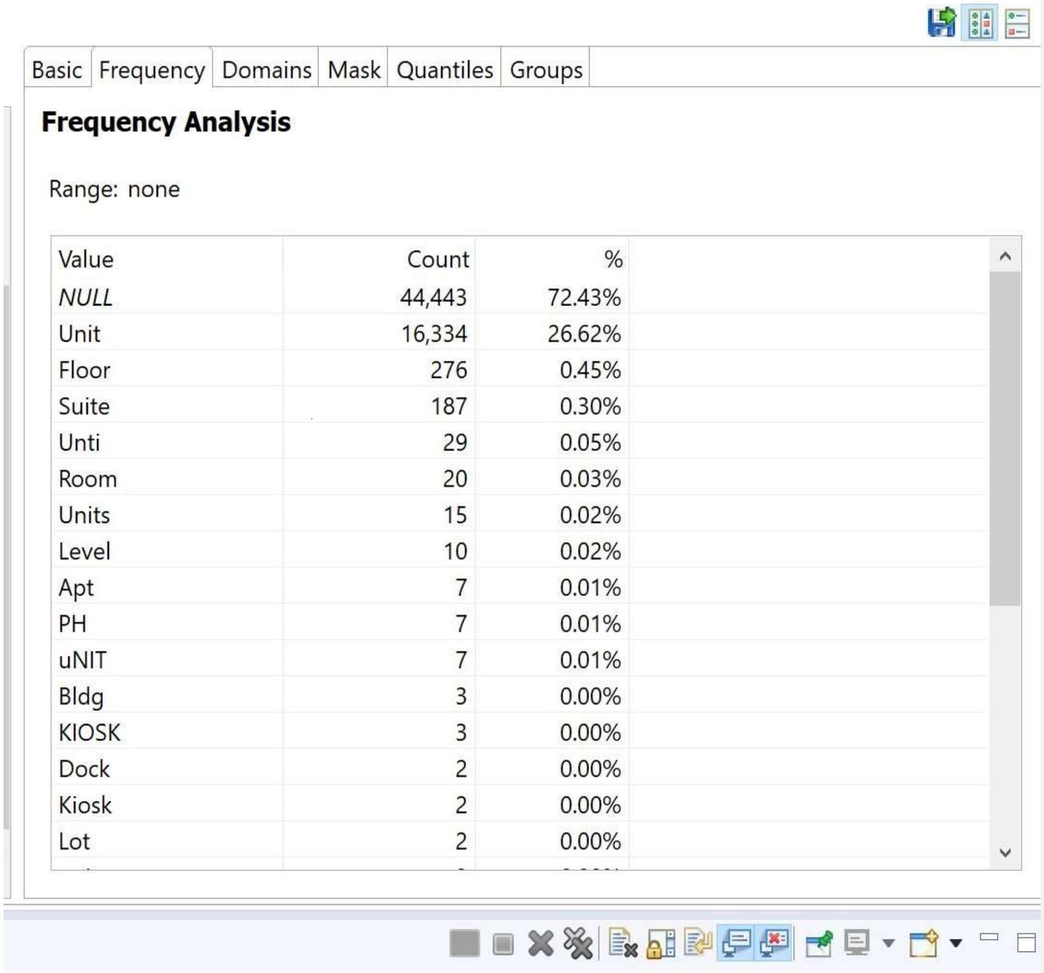


LicenceRSN
1
3084939 Count
3097759 Count
3187938 Count
3193480 Count
3219860 Count
3221566 Count
3225267 Count
3225501 Count
3225502 Count
3225541 Count
3225542 Count
3225780 Count
3225784 Count
3225793 Count
3225804 Count
3225823 Count
3225867 Count
3225888 Count
3225891 Count
3226239 Count
3226240 Count
3226241 Count
3226242 Count
3226243 Count
3226244 Count
3226245 Count
3226246 Count

If you click on count slicer on 1 it shows all uniq licenceRSN and if you click on same slicer on 2 shows dublicated licenceRSN,if you click on which licenceRSN it will shows that record
with BussinessName slicer can select a bussiness and see the LicenceRSN for that bussiness.

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.

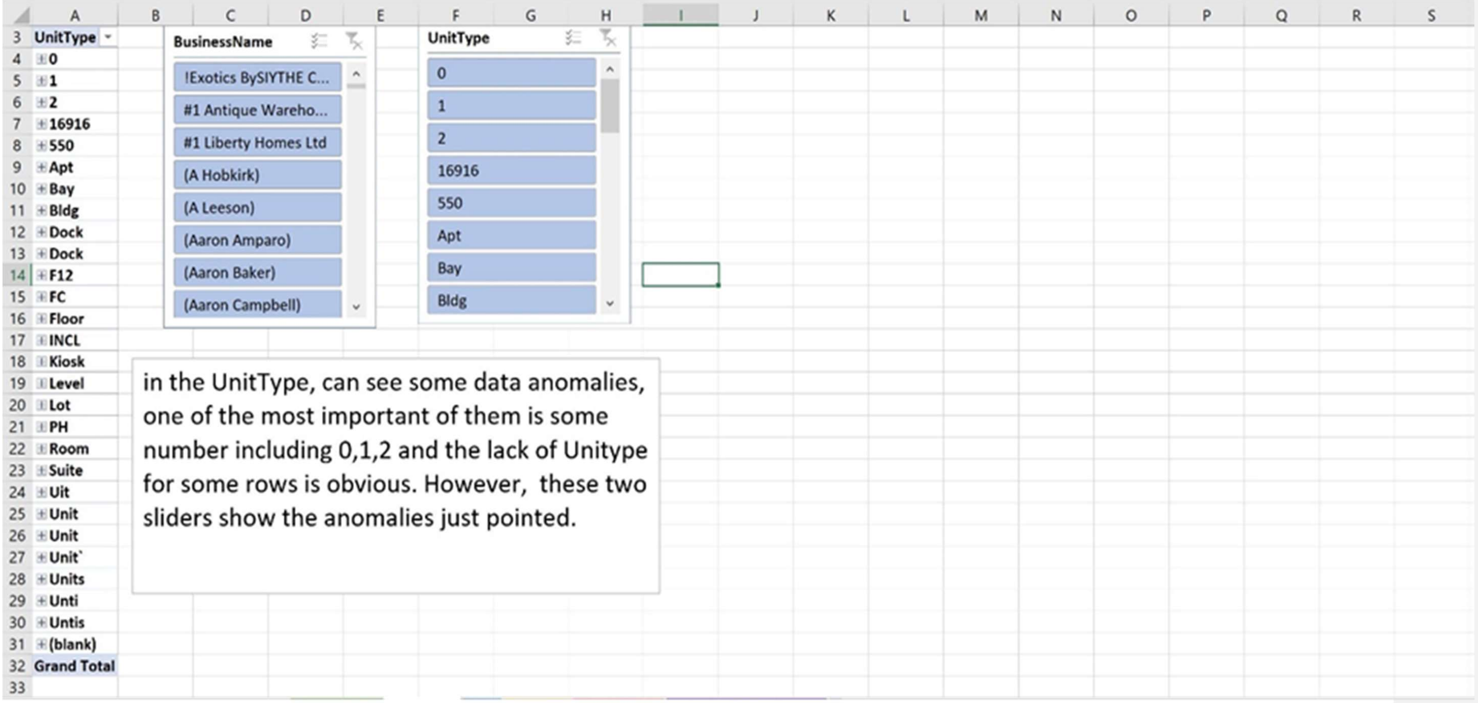


Frequency Analysis

Range: none

Value	Count	%
NULL	44,443	72.43%
Unit	16,334	26.62%
Floor	276	0.45%
Suite	187	0.30%
Unti	29	0.05%
Room	20	0.03%
Units	15	0.02%
Level	10	0.02%
Apt	7	0.01%
PH	7	0.01%
uNIT	7	0.01%
Bldg	3	0.00%
KIOSK	3	0.00%
Dock	2	0.00%
Kiosk	2	0.00%
Lot	2	0.00%

]

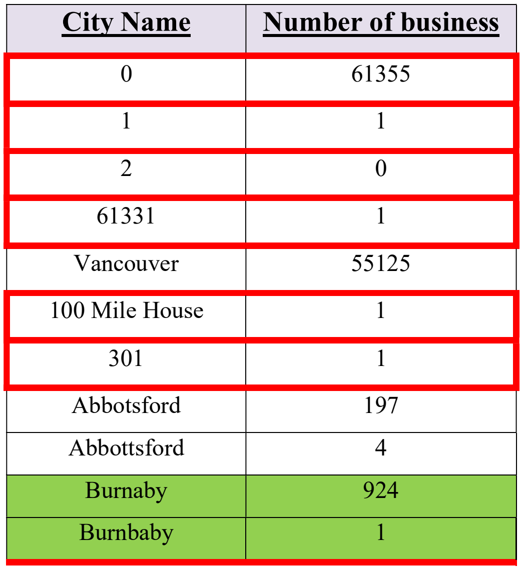


UnitType
0
1
2
16916
550
Apt
Bay
Bldg
Dock
Dock
F12
FC
Floor
INCL
Kiosk
Level
Lot
PH
Room
Suite
Uit
Unit
Unit
Unit`
Units
Unti
Untis
(blank)
Grand Total

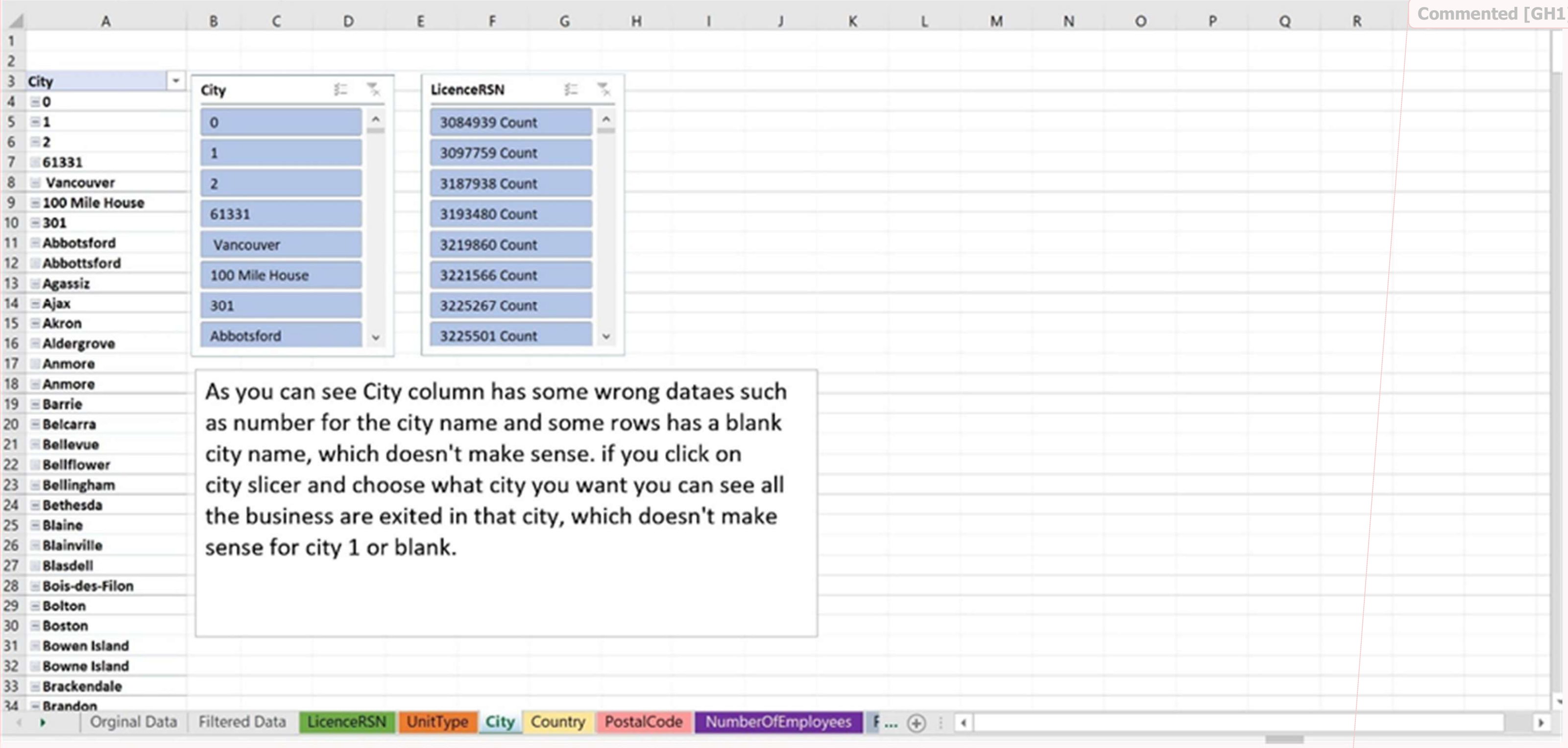
in the UnitType, can see some data anomalies, one of the most important of them is some number including 0,1,2 and the lack of Unitype for some rows is obvious. However, these two sliders show the anomalies just pointed.

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).



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



City
0
1
2
61331
Vancouver
100 Mile House
301
Abbotsford
Abbottsford
Agassiz
Ajax
Akron
Aldergrove
Anmore
Anmore
Barrie
Belcarra
Bellevue
Bellflower
Bellingham
Bethesda
Blaine
Blainville
Blasdell
Bois-des-Filon
Bolton
Boston
Bowen Island
Bowne Island
Brackendale
Brandon

City	LicenceRSN
0	3084939 Count
1	3097759 Count
2	3187938 Count
61331	3193480 Count
Vancouver	3219860 Count
100 Mile House	3221566 Count
301	3225267 Count
Abbotsford	3225501 Count

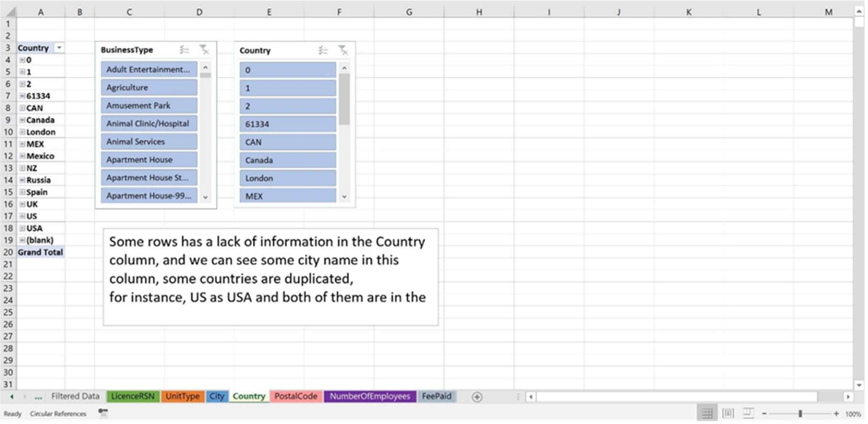
As you can see City column has some wrong dataes such as number for the city name and some rows has a blank city name, which doesn't make sense. if you click on city slicer and choose what city you want you can see all the business are exited in that city, which doesn't make sense for city 1 or blank.

Commented [GH1]: Okay

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.

Country	**Number of Business**
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



Country
0
1
2
61334
CAN
Canada
London
MEX
Mexico
NZ
Russia
Spain
UK
US
USA
(blank)
Grand Total

BusinessType

- Adult Entertainment...
- Agriculture
- Amusement Park
- Animal Clinic/Hospital
- Animal Services
- Apartment House
- Apartment House St...
- Apartment House-99...

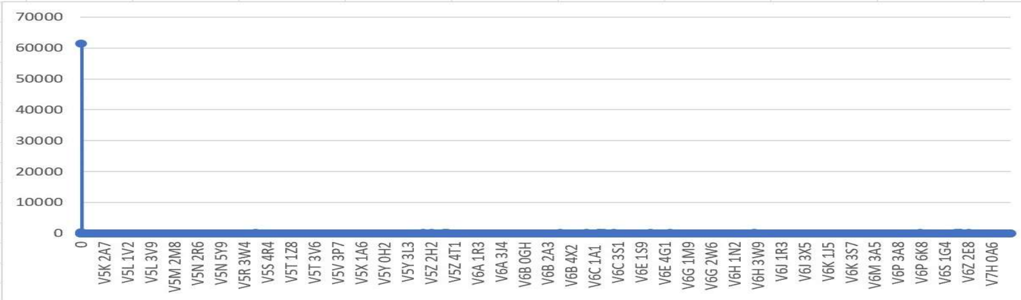
Country

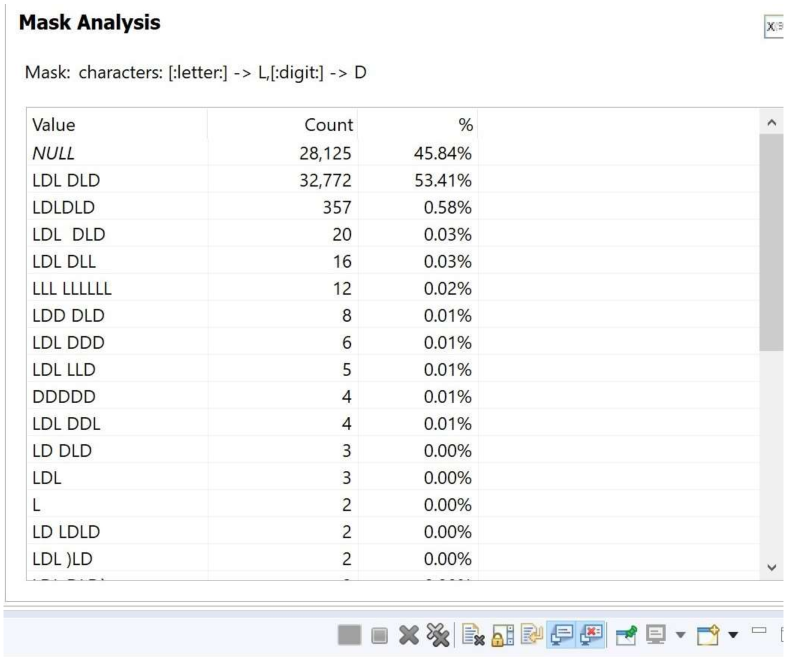
- 0
- 1
- 2
- 61334
- CAN
- Canada
- London
- MEX

Some rows has a lack of information in the Country column, and we can see some city name in this column, some countries are duplicated,
for instance, US as USA and both of them are in the

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.





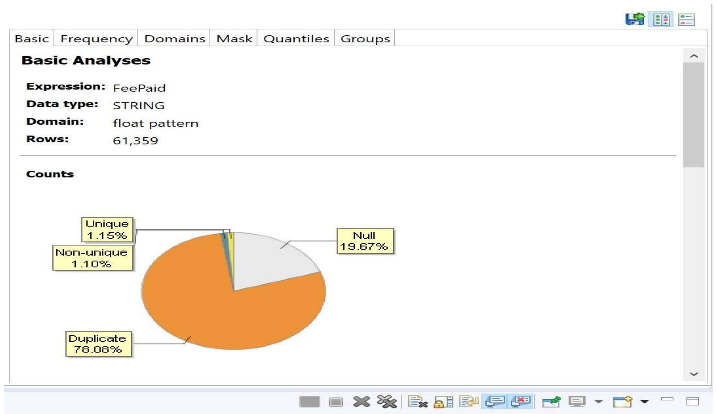
Mask Analysis

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

Value	Count	%
NULL	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

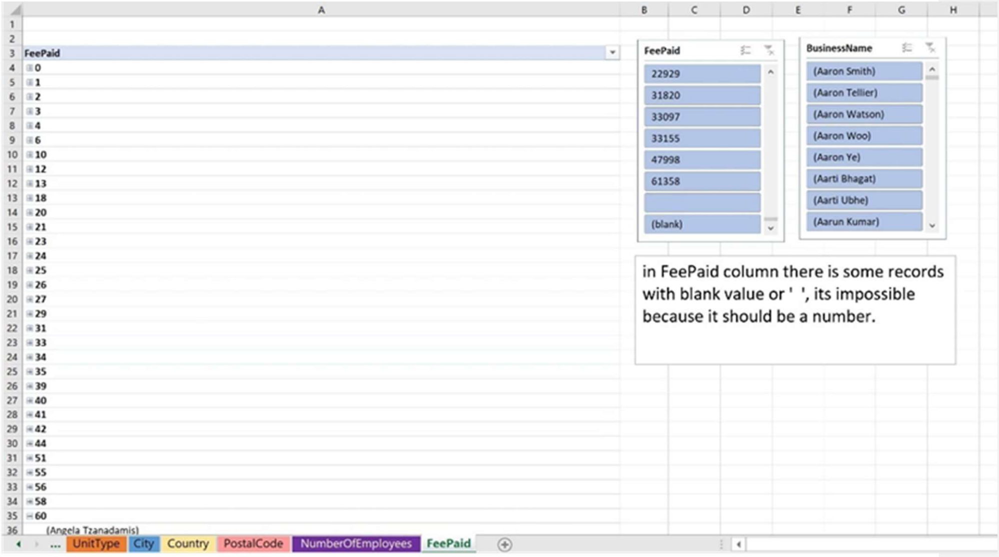
Initial analysis of the FeePaid column demonstrated that there exist 19.67% cells is Null. This might make some categorizing and analyzing issue in the next step of data analyzing.



Basic Analyses

Expression: FeePaid
Data type: STRING
Domain: float pattern
Rows: 61,359

Counts



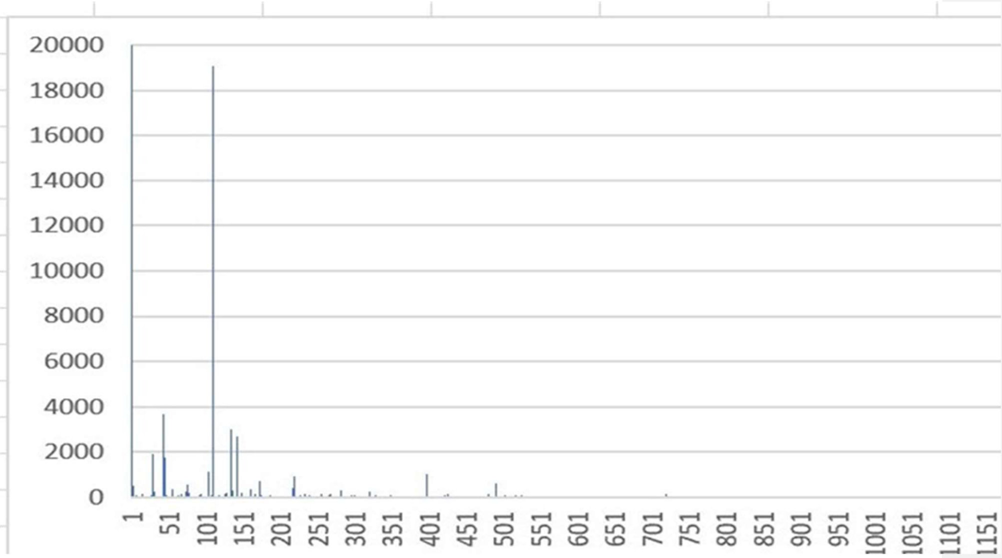
FeePaid
0
1
2
3
4
6
10
12
13
18
20
21
23
24
25
26
27
29
31
33
34
35
39
40
41
42
44
51
55
56
58
60
(Angela Tzanadamis)

FeePaid
22929
31820
33097
33155
47998
61358

(blank)

BusinessName
(Aaron Smith)
(Aaron Tellier)
(Aaron Watson)
(Aaron Woo)
(Aaron Ye)
(Aarti Bhagat)
(Aarti Ubhe)
(Aarun Kumar)

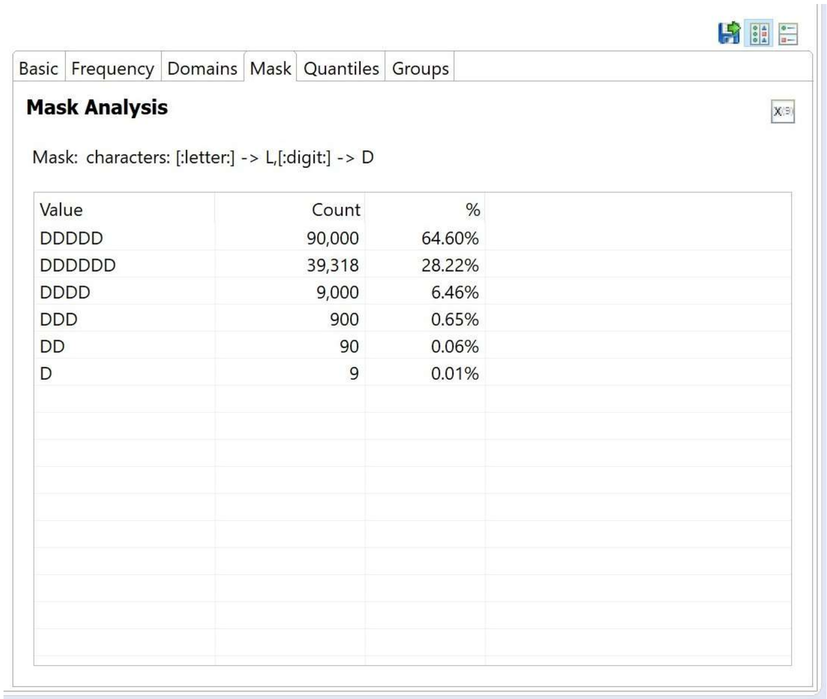
in FeePaid column there is some records with blank value or ' ', its impossible because it should be a number.



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**.



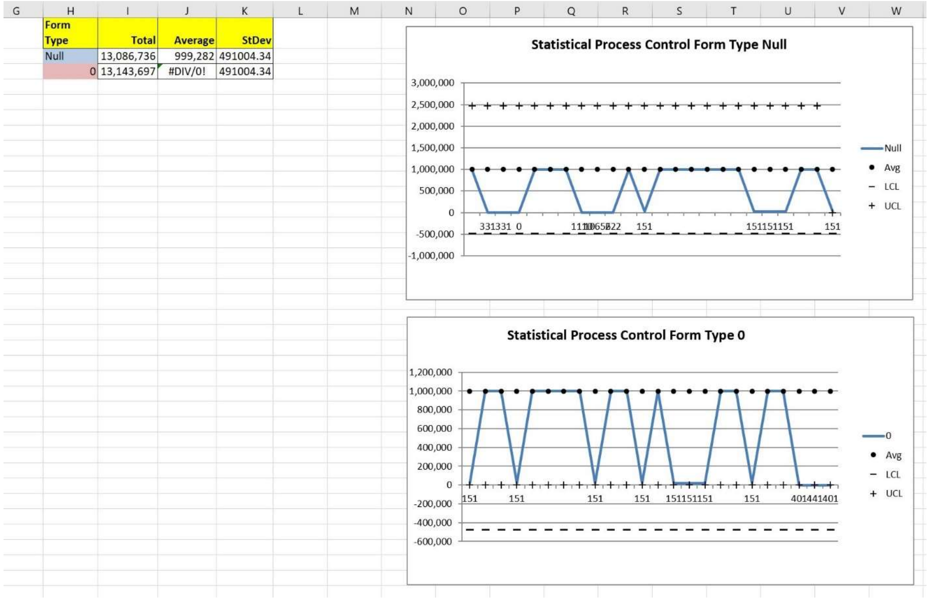
Mask Analysis

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

Value	Count	%
DDDDD	90,000	64.60%
DDDDDD	39,318	28.22%
DDDD	9,000	6.46%
DDD	900	0.65%
DD	90	0.06%
D	9	0.01%

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.



Form Type	Total	Average	StDev
Null	13,086,736	999,282	491004.34
0	13,143,697	#DIV/0!	491004.34

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**.