Project Specs

COMP 3839

Shane Calado, A00817064

Final Version

Contents

Project Specs....1
Overview of Project....2
Initial Assessment....3
BusinessName: NULL values....3
City: Inconstancies/misspelling in data....3
Country: 3 letter naming standard not followed....4
FeePaid: Null values....4
PostalCode: Inconstancies in mask....5
Province: Inconstancies in mask and values....5
SME Review of Initial Assessment....6
BusinessName: NULL values....6
City: Inconstancies/misspelling in data....6
Country: 3 letter naming standard not followed....6
FeePaid: Null values....6
PostalCode: Inconstancies in mask....6
Province: Inconstancies in mask and values....6
Further Research for the SME....7
SME Review of Further Research....9
SME Suggests Some DQ Rules....10
BusinessName: Required field....10
Province: Mask to force 2 letter naming standard....10
Country: Mask to force 3 letter naming standard....10
FeePaid: Report unpaid entries....10

Overview of Project

This report is an assessment of the 2018 Vancouver Business Licence data set. The dataset contains information on all applicants to the business licence program such as the name of the business, location data, the type of business, size of the business (number of employees), as well as the status of the licence application, data on the fees paid by the business, and issue and expiry dates for those businesses with valid licences.

Column	Data-type	Description
BusinessName	STRING	Name of Business
BusinessType	STRING	Type of Business
BusinessSubType	STRING	
BusinessTradeName	STRING	Trade of the business (if applicable)
City	STRING	City the business is located in
Country	STRING	Country the business is located in
ExpiredDate	DATETIME	Date the business licence expires.
ExtractDate	DAY	
FeePaid	FLOAT	Total fees paid by the applicant.
House	STRING	House the business is located in
IssuedDate	DATETIME	Date the business licence was issued.
Latitude	STRING	Latitude the business is located at.
LicenceNumber	STRING	Licence number issued to the business.
LicenceRevisionNumber	STRING	Revision to the licence number (if applicable)
LicenceRSN	INTEGER	LicenceRSN the business is located in
LocalArea	STRING	Local Area the business is located in.
Longitude	STRING	Longitude the business is located at.
NumberOfEmployees	STRING	Number of employees the business has
PostalCode	STRING	Postal code the business is located at
Province	STRING	Province the business is located in
Status	STRING	Current licences status of the business.
Street	STRING	Street the business is located at.
Unit	STRING	Unit the business is located in.
UnitType	STRING	Type of unit the business is located in.

We will be profiling all the fields within the dataset, and making note of any outliers within the data. We will then present our findings to Bob Barker, the manager of the Vancouver Licence Office and requesting insight from him on which findings indicate a potential issue, or require further research.

Initial Assessment

BusinessName: NULL values

- There are 3,753 records that have a BusinessName of "NULL".
- 3250 of those records also have a Status of "Issued".

Type	Count	%
Null	3,753	5.67%
Non-null	62,427	94.33%
Duplicate	8,721	13.18%
Distinct	53,706	81.15%
Non-unique	5,118	7.73%
Unique	48,588	73.42%

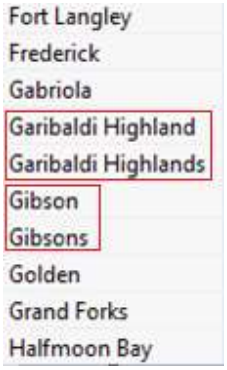
Status of licences with NULL BusinessName

Issued	3250
Cancelled	273
Pending	158
Gone out of Business	51
Inactive	21

City: Inconstancies/misspelling in data

- There are multiple misspelled entries in the data.

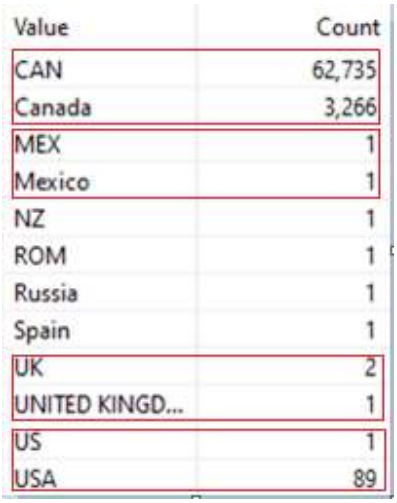
Type	Count	%
Null	86	0.13%
Non-null	66,094	99.87%
Duplicate	65,828	99.47%
Distinct	266	0.40%
Non-unique	110	0.17%
Unique	156	0.24%



Country: 3 letter naming standard not followed

- The entries don't seem to follow a naming standard. For example, "Canada" is also represented as "CAN", "Mexico" as "MEX", and USA as "US" and "United States".

Value	Count
CAN	62,735
Canada	3,266
MEX	1
Mexico	1
NZ	1
ROM	1
Russia	1
Spain	1
UK	2
UNITED KINGD...	1
US	1
USA	89



Value	Count
CAN	62,735
Canada	3,266
MEX	1
Mexico	1
NZ	1
ROM	1
Russia	1
Spain	1
UK	2
UNITED KINGD...	1
US	1
USA	89

FeePaid: Null values

- There are 154 records where FeePaid is Null and the status is Issued.

Type	Count	%
Null	9,298	14.05%
Non-null	56,882	85.95%
Duplicate	55,361	83.65%
Distinct	1,521	2.30%
Non-unique	760	1.15%
Unique	761	1.15%

Status of Licences with NULL FeePaid	
Pending	5412
Gone out of Business	2876
Cancelled	835
Issued	154
Inactive	21

PostalCode: Inconstancies in mask

- Inconstancies in mask resulting in incorrect postal codes, all entries should follow a set format.
- The format either in format LDL DLD(Canada) or DDDDD(USA).

Value	Count
NULL	29,547
LDL DLD	36,081
LDLDLD	384
LDL DLD	76
LDL DLL	17
LLL LLLLLL	10
LDD DLD	8
LDL DDD	8
DDDDD	6
LDL LLD	5
LDL DDL	4
L/L	3
LD DLD	3
LDL	3

L	2
LD LDLD	2
LDL)LD	2
LDL DLD`	2
LDLL DLD	2
LLL DLD	2
DDDDDD	1
DDDDDDDDDD	1
DDL DLD	1
DLLDLD	1
LD DLL	1
LD: DLD	1
LD& DLD	1
LDL DD	1
LDL DLDDDDDD	1
LDLDDD	1

Province: Inconstancies in mask and values

- There are inconstancies in the mask that cause redundancies in the data, for example “BC” is also represented as “British Columbia” / “AB” as “Alberta” etc…

Type	Count	%
Null	94	0.14%
Non-null	66,086	99.86%
Duplicate	66,048	99.80%
Distinct	38	0.06%
Non-unique	23	0.03%
Unique	15	0.02%

Value	Count
NULL	94
`	1
YT	1
WA	12
VA	1
TX	2
SK	2
SC	1
QC	16
PQ	5
PA	1
On	3
OR	1
ON	110
OH	4
NY	13

SME Review of Initial Assessment

BusinessName: NULL values

- High priority.
- Unsure how businesses are able to be issued a licence without registering a name.
- Please look into this further.

City: Inconstancies/misspelling in data

- Low priority
- Seems like more of a housecleaning issue, should be done eventually.

Country: 3 letter naming standard not followed

- Low priority
- This makes it more difficult to make quick observations of the data, but beyond that does seem to have a high impact on the business.

FeePaid: Null values

- High priority
- We need to ensure that all businesses with an Issued licence have paid the appropriate fees, can a list of all missing fees be compiled?

PostalCode: Inconstancies in mask

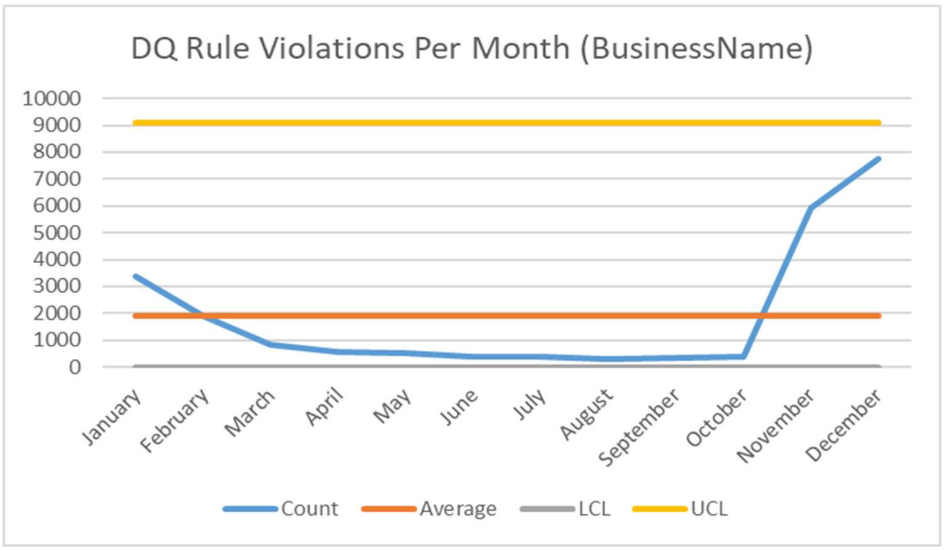
- Medium priority
- Incorrect postal code will not allow us to contact or invoice businesses by mail.

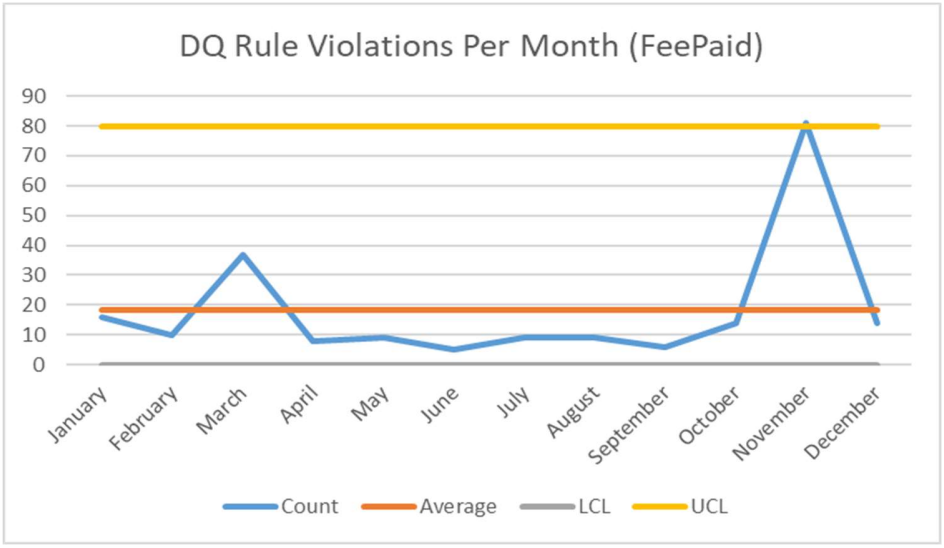
Province: Inconstancies in mask and values

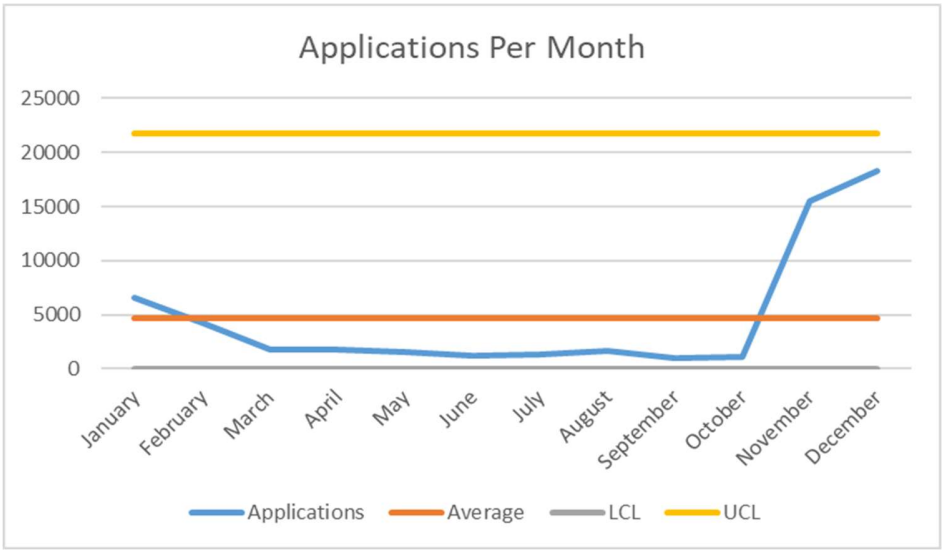
- Low priority
- Like the Country field, this makes it more difficult to make quick observations of the data but beyond that its effects are limited.

Further Research for the SME

- The amount of Businesses that have Issued licence and either no name or an unpaid fee, appears to be highest between the months of October to January, although both follow the same trend as the amount of applications per month so it may not be abnormal.
- The trend of FeePaid rule violations varies from the amount of applications, but this could also be due to smaller counts of the violation causing more fluctuation in the graph.
- These charts were calculated by comparing the dataset to 2 rules (BusinessName = NULL and Status = "Issued"/ Status = "Issued" and FeePaid <0 or NULL). The resulting counts of rule violations were then sorted by the month in which the licence was issued.


DQ Rule Violations Per Month (BusinessName)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
January
February
March
April
May
June
July
August
September
October
November
December
Count
Average
LCL
UCL


DQ Rule Violations Per Month (FeePaid)
90
80
70
60
50
40
30
20
10
0
January
February
March
April
May
June
July
August
September
October
November
December
Count
Average
LCL
UCL


Applications Per Month
25000
20000
15000
10000
5000
0
January
February
March
April
May
June
July
August
September
October
November
December
Applications
Average
LCL
UCL

SME Review of Further Research

- Although there seems to be an abnormally high concentration of business rule violations in the months of October to January, the SPC chart shows that it is still in line with the volume of applications coming in at that time. SME initially had concerns over spike in rule violations but after reviewing the SPC chart, has decided that the results are not out of line (All results are within the control limits).

SME Suggests Some DQ Rules

BusinessName: Required field

- Flag and reject any entries that come through the ETL without a BusinessName as a rule violation.

Province: Mask to force 2 letter naming standard

- Any entries that come through the ETL not following the 2 letter naming standard will be flagged as a rule violation.
- Any entries that can be automatically corrected to the standard (ex. “British Columbia” to “BC”) will be changed but still flagged as a rule violation.

Country: Mask to force 3 letter naming standard

- Any entries that come through the ETL not following the 3 letter naming standard will be flagged as a rule violation.
- Any entries that can be automatically corrected to the standard (ex. “Canada” to “CAN”) will be changed but still flagged as a rule violation.

FeePaid: Report unpaid entries

- Flag any entries that come through the ETL that have an “Issued” licence and a FeePaid < 0.