COMP3839   
Azarm Piran  
A01195657

Contents

[Overview of Project 1](#_Toc58185344)

[Initial Assessment 2](#_Toc58185345)

[1) LicenceRSN : Duplicates 2](#_Toc58185346)

[2) BusinessName: Nulls 2](#_Toc58185347)

[3) Province: Invalid Entries 3](#_Toc58185348)

[4) Postal Code: Invalid Format 5](#_Toc58185349)

[5) City: Mis-spellings 6](#_Toc58185350)

[6) FeePaid: Null and Zero Values 8](#_Toc58185351)

[SME Review of Initial Assessment 12](#_Toc58185352)

[1) LicenceRSN : Duplicates 12](#_Toc58185353)

[2) BusinessName: Nulls 12](#_Toc58185354)

[4) Postal Code: Invalid Format 12](#_Toc58185355)

[5) City: Mis-spellings 12](#_Toc58185356)

[6) FeePaid: Null and Zero Values 13](#_Toc58185357)

[Further Research for the SME 13](#_Toc58185358)

[SME Review of Further Research 14](#_Toc58185359)

[SME Suggests Some DQ Rules 14](#_Toc58185360)

[1) LicenceRSN : Duplicates 14](#_Toc58185361)

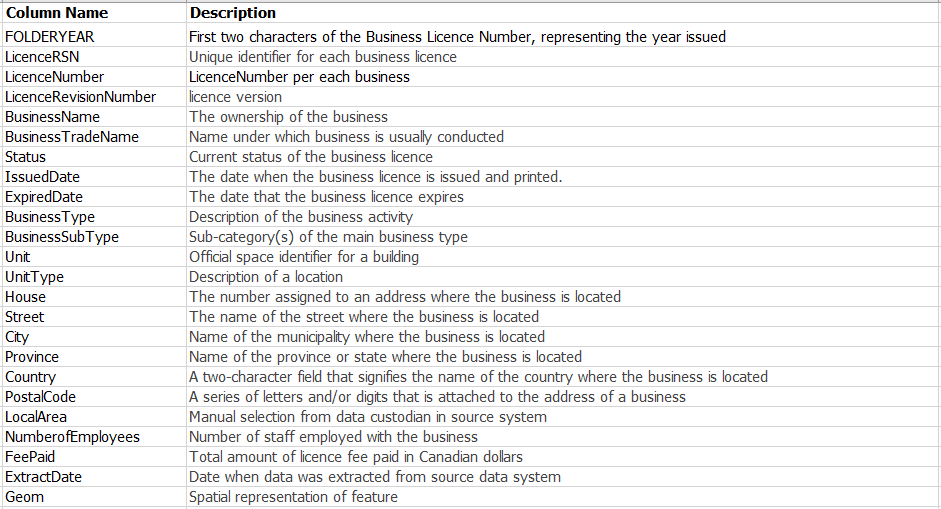
[2) BusinessName: Nulls 14](#_Toc58185362)

[3) Province: Invalid Entries 15](#_Toc58185363)

[4) Postal Code: Invalid Format 15](#_Toc58185364)

[5) City: Mis-spellings 15](#_Toc58185365)

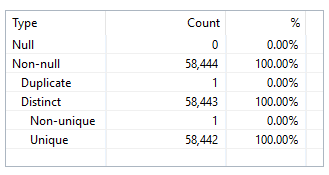
[6) FeePaid: Null and Zero Values 15](#_Toc58185366)

Overview of Project  
The dataset in this project contains 58444 records of Vancouver Business Licence in 2019.  
We can see a brief description of each column of data:  
  
We can find the data source in the following address:  
 [Vancouver Business Licence 2019](https://opendata.vancouver.ca/explore/dataset/business-licences/information/?disjunctive.status&disjunctive.businesssubtype&refine.issueddate=2019)

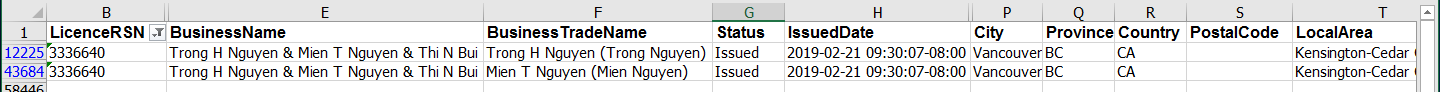
My goal will be looking for issues in data, showing a couple of samples per each issue and making suggestions to solve these anomalies.   
The first part of the project will be indicating each issue with a brief description and context table and sample of issues in real data.   
Initial assessment by SME is in the second part of the project and it is mainly about whether she approves the issue or not and the priority for each issue. Azin Piran has been the head of licencing for more than 15 years at the City of Vancouver and she will be guiding me as the SME.  
In the third part of the project, we will focus on one of the issues and create an SPC chart for that.  
The last part of the project is about the SME’s suggestions for each issue in the dataset.

# Initial Assessment

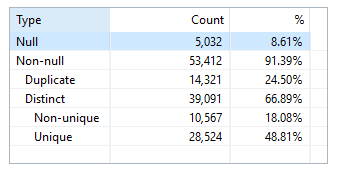
1) LicenceRSN : Duplicates   
LicenceRSN is a primary key and a unique identifier in this dataset. Based on the screenshot below we can see that there are 1 Duplicate record and 1 Non-unique record in our primary key. Duplication in a unique identifier column causes some difficulties in identification of records.



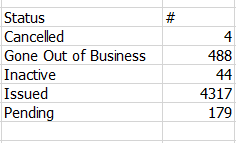
Here is the duplication issue in LicemceRSN column in data, as we can see we have two records with the same ID.



2) BusinessName: Nulls  
The business name is the ownership of business so it is such a piece of important information about a business. We have 5032 records which are 8.61% of data with no business name.   
8.61% percentage of data is null which is a lot especially because most of them are businesses with status = issued.

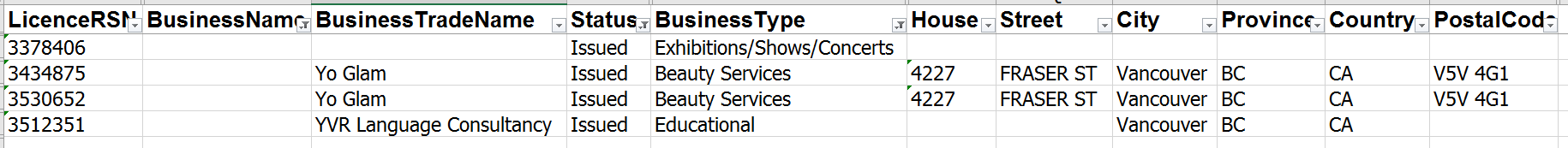


Here is the number of records with BusinessName = Null per each status. A large portion of nulls falls into the issued status part.

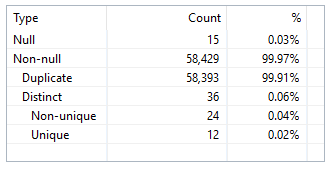


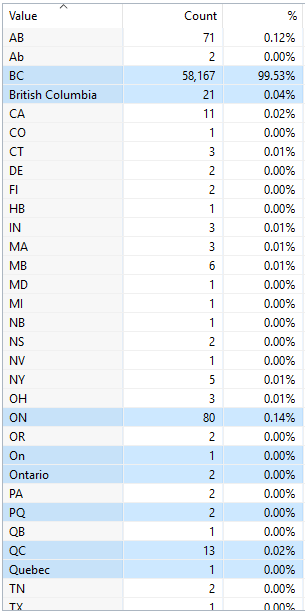
In this screenshot, we will focus on the count of different business types with BusinessName = Null and Status = Issued:  
As we can see, a substantial portion of BusinessName = Null and Status = Issued belongs to Short-Term Rental business type.  


Here we have more detail for 2 records in BusinessType = “Beauty Services” and 1 record for “Educational” as well as a one record for “Exhibitions/Shows/Concerts”.

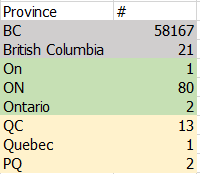


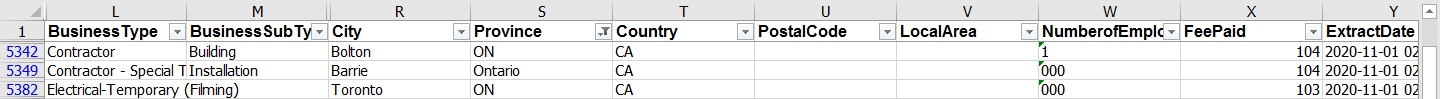
3) Province: Invalid Entries  
We are using two different standards for province column. Two-digit standard and in some cases the whole name of the province such as BC and British Columbia which indicate the same province but we used two different standards in data. We have the same problem for ON and Ontario, Quebec and QC and so on.

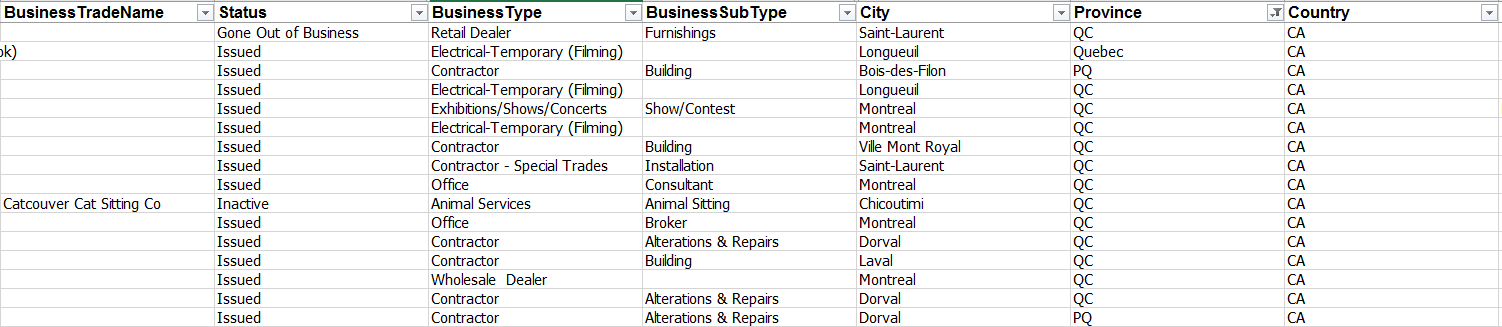


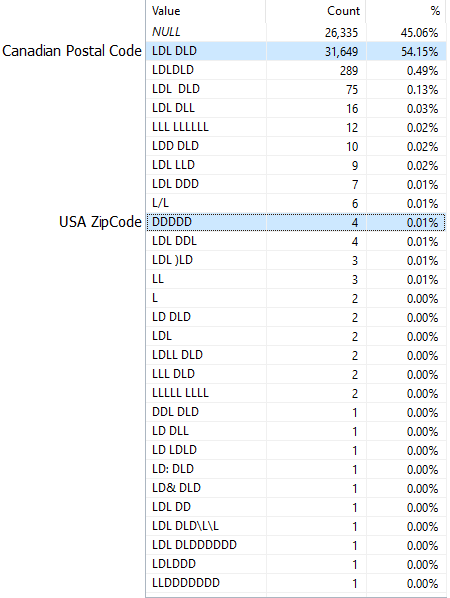
As we can see in the frequency tab there are 2 different formats for some of the provinces.  


As we can see two letter format for province name is used more than the complete name format.



Here is the example for Ontario and ON, two different standards.  


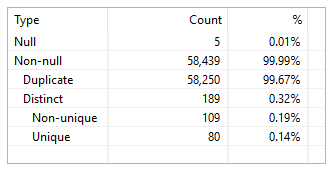
Here is another example of different formats for the province column, QC and Quebec.  


4) Postal Code: Invalid Format  
Canadian postal code format is: LDL DLD  
The second row in the picture below shows the correct format which is 31649 records.  
4th rows show the number of 75 records with LDL DLD format.  
USA Zipcode format is 5 digits. There are only 4 records with this format.  
I believe this is an issue when we are trying to find postal codes within a specific range or specific format.   
For example, all the postal code in British Columbia starts with letter V. the second character is used to seprate rural and urban regions.   
I am looking for all the rural postal codes in BC which starts with V0.   
The problem here is:   
I have different formats for BC’s rural postal codes like this V 0A 1B2, V0A 1 B2 and V0 A 1B2.  
In this case, it is impossible to find all the BC’s rural postal codes with the same format and there is chance that we miss some of the formats in our search and then the results will not be accurate.  
  


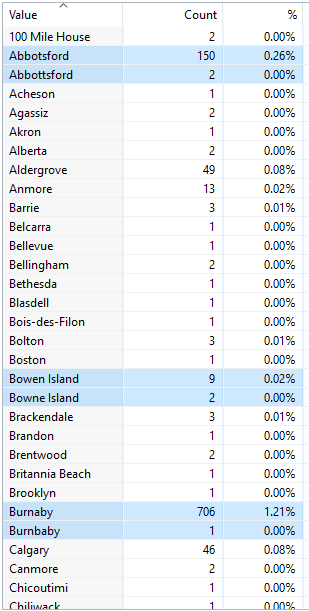
5) City: Mis-spellings  
In the city column, we have several mis-spellings which cause two different groups of data for the same city such as Burnaby and Burnbaby.

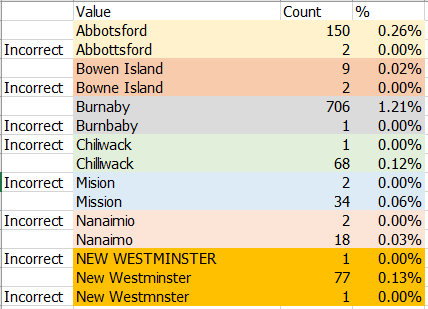
In case we are looking for all the businesses located in the city of Burnaby, we are going to miss some of the businesses due to the mis-spelling issue because the value of city for some records is **Burnbaby** not Burnaby. This is an issue because the result will not be correct. We will miss some part of data in our reports and analysis.

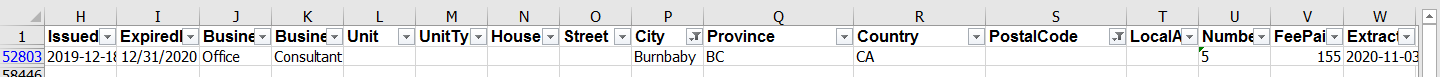
Here is the context table:

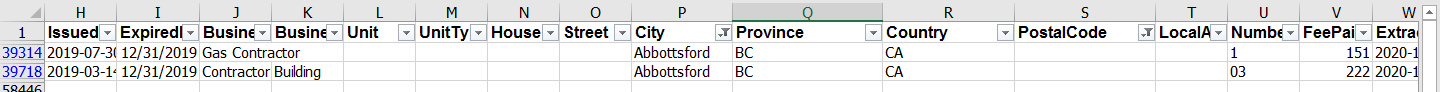


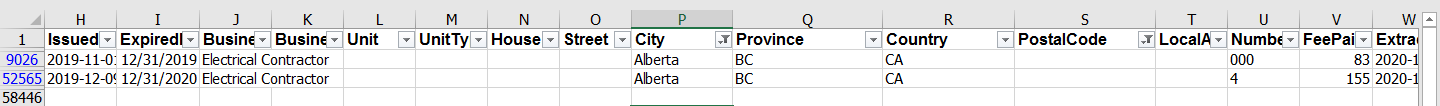
The frequency table for this column is as follows:



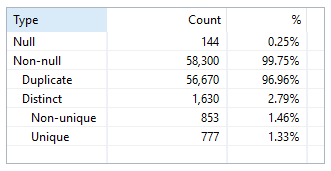
In this table, we can see a couple of mis-spelling cases in data with the number of occurrence and the percentage of whole data:  


Here is the mis-spelling in the city column for Burnaby in data.  


Here is the mis-spelling in the city column for Abbotsford in data.   


Alberta is not the name of the city, it is the province name.  


6) FeePaid: Null and Zero Values  
This column represents the total amount of licence fee paid in Canadian dollars which is such important information. We have 144 nulls in this column.



This table shows the status for 144 records with a null value in FeePaind column. There are 109 records with Status = Issued value.

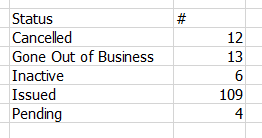


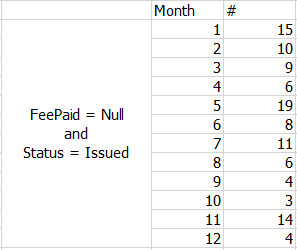
Table below represents more analysis on 109 records (FeePaid = null and Status = Issued) based on each month of year.   
5th month of year has the highest amount.  


Table below represents more analysis on 109 records (FeePaid = null and Status = Issued) based on each city.   
Vancouver with 101 records has the highest amount.

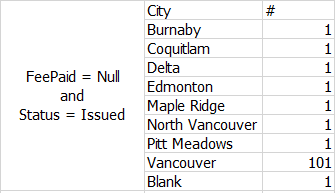


Table below represents more analysis on 109 records (FeePaid = null and Status = Issued) based on BusinessType.   
“Temp Liquor Licence Amendment” with 8 records has the highest amount.  
I think businessType is not a good criterion for further analysis since there are small amounts for each type.

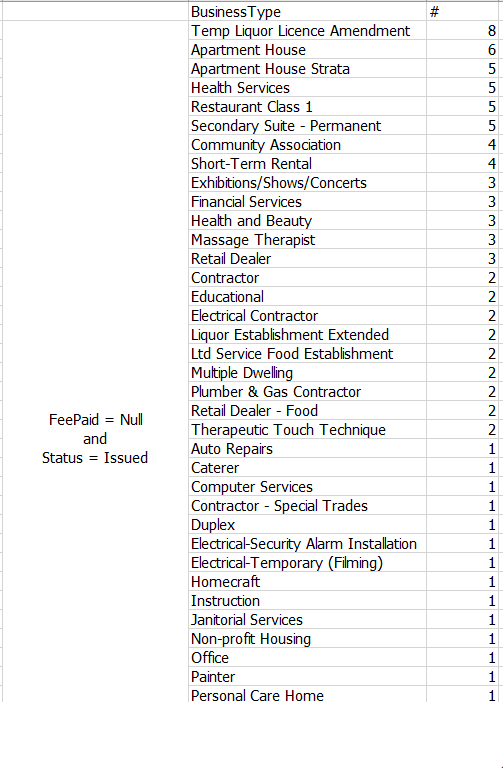
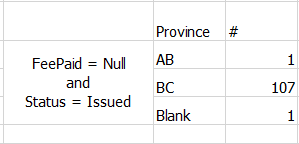


Table below represents more analysis on 109 records (FeePaid = null and Status = Issued) based on each province. BC with 107 records has the highest amount.



Province and city are appropriate criterions for further analysis.

# SME Review of Initial Assessment

1) LicenceRSN : Duplicates  
Priority = High  
I affirm this issue, based on my research there are two records with the same LicenceRSN number. LicenceRSN is a primary key and unique identifier and this issue could be a big problem in identifying businesses.

2) BusinessName: Nulls  
Priority = High  
There are 4317 issued businesses with a null value. A business name is an important part of each business. Since a big portion of null values belongs to issued business the priority to fix this column is high.

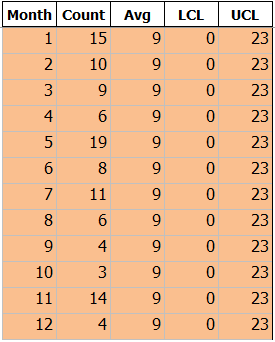
‎3) Province: Invalid Entries  
Priority = Low  
Yes, there is no standard format for provinces in this column and it needs to get fixed but the priority is low since we still can handle it.

4) Postal Code: Invalid Format  
Priority = Low  
Yes, I noticed that we do not use the standard format for Canadian postal codes and in some cases USA ZipCodes. The priority for this issue is low since more than half a portion of data 54.15% follows the correct format.

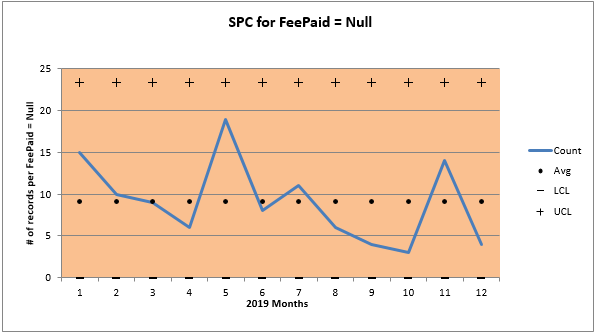
5) City: Mis-spellings  
Priority = Low  
Fixing the mis-spelling issue in this column has low priority since the wrong dictations are less than 1% of the whole data. This will cause several groups for the same city and we will have problems in data classifications in future.

6) FeePaid: Null and Zero Values  
Priority = Low  
The percentage of businesses which have not paid their fee paid is less than 1 percentage of whole data. Plus, not all of the businesses with FeePaid = null are issued, which makes it even less than 0.25 %.  
Since this is an important information, please undertake further research for this issue to see whether it is under contro or not.

Further Research for the SME  
FeePaid: Null and Zero Values  
Based on the SME request, in this part, we did some further research for FeePaid column.  
We have 144 records which have not paid their fees.  
Out of 144 records, there are 109 records with status = issued which have not paid their fees. This table represents the number of 109 FeePaid = Null and Status = Issued records in each month of the year. The highest number of records belongs to May.



Now we have this table, we create an SPC chart, here is the result:



# SME Review of Further Research

Thanks for the deep analysis, as we can see in SPC most amount of disorder occurred around 5th month of the year. The SPC is under control. We do not need further research.

# SME Suggests Some DQ Rules

1) LicenceRSN : Duplicates  
It is a good idea to set a unique value to each record in LicenceRSN column to avoid duplications in future and identification.  
My suggestion is to set an alert when duplicate LicenceRSNs are found in the table.‎

2) BusinessName: Nulls  
I recommend to update the software to not allow any null values for this column when the status is issued. Status = Issued means the completion of the process and business needs a name in this level.  
My suggestion is to set an alert when Status=Issued and BusinessName=Null‎

3) Province: Invalid Entries  
It is better to have one indicator for each Province. I recommend to enforce the abbreviation format of provinces like BC instead of British Columbia.  
My suggestion is to set an alert when Status=Issued and Province is not 2 characters‎ format.

4) Postal Code: Invalid Format  
It would be really helpful if we can have the same format for all postalCodes in data and standardize the postal code based on each countries postal code ‎format.‎  
My suggestion is to set an alert when PostalCode does not match the format expected for the Country‎.

5) City: Mis-spellings  
We can have a dropdown box to choose an appropriate city to avoid mis-spelling.  
My suggestion is to set an alert when City is not found in the Lookup city table.‎

6) FeePaid: Null and Zero Values  
We are missing important information in this column. My suggestion is to set an alert when Status=Issued and FeePaid=Null‎

‎