**Contents**

[**Overview of Project** 2](#_Toc182578114)

[**Initial Assessment** 3](#_Toc182578115)

[PostalCode: Nulls 3](#_Toc182578116)

[PostalCode: Format 4](#_Toc182578117)

[LicenceRSN: Duplicate 6](#_Toc182578118)

[Province: Format 7](#_Toc182578119)

[Status: Invalid input 8](#_Toc182578120)

[**SME Review of Initial Assessment** 9](#_Toc182578121)

[PostalCode: Null 9](#_Toc182578122)

[PostalCode: Format 9](#_Toc182578123)

[LicenseRSN: Duplicate 9](#_Toc182578124)

[Province: Format 10](#_Toc182578125)

[Status: Invalid input 10](#_Toc182578126)

[**Further Research for SME** 10](#_Toc182578127)

[LicenseRSN: Duplicate 10](#_Toc182578128)

[**SME Review of Further Research** 12](#_Toc182578129)

[**SME Suggestions for DQ Rules** 12](#_Toc182578130)

[PostalCode: Format and Null 12](#_Toc182578131)

[LicenseRSN: Duplicate 12](#_Toc182578132)

[Province: Format 13](#_Toc182578133)

[Status: Invalid input 13](#_Toc182578134)

# **Overview of Project**

The dataset that we are profiling is by the City of Vancouver which records business license information for the year 2020. Each business in Vancouver is required to have a valid business license, and this dataset keeps track of metrics related to issuing the license such as dates, business name, license number, fees. It also houses information related to the business itself like address, business type, and number of employees.

* LicenceRSN – this is the primary key that only consists of a number for the table that is generated.
* LicenceNumber – is a set 9 character field consisting of two sets of numbers. First 2 digits are the year of issue business license and the next 6 are generated.
* LicenceRevisionNumber – is 2 digits long and represents the version of the license.
* BusinessName – this is a string that represents ownership of the business.
* BusinessTradeName – a string representing name of business.
* Status – it is the current of the license for the business ranging from issue, inactive, cancelled, etc.
* IssuedDate – date and time of license being issued
* ExpiredDate – date of license expiring
* BusinessType and BusinessSubType – the type of activity associated with the business which can contain a subtype. This is stored as a string
* Unit, UnitType, House, Street, City, Province, Country, PostalCode, LocalArea – these are all sub components of a complete address stored at string.
* NumberofEmployees – a number that represents the amount of people working at that business.
* FeePaid – the amount of money business paid for the license in Canadian dollars
* ExtractDate – date of data being extracted from system

The Subject Matter Expert for this dataset understands the process of issuing business licenses, housing data for those licenses, as well as an increase in technical understanding.

Data profiling will be done using Ataccama Data Quality Analyzer. Through this tool, we are checking to see if there are any anomalies that exist at the column level. This can include identifying columns with high null count, frequent outliers, the range and format of each record, or columns not being used for its stated purpose.

# **Initial Assessment**

## PostalCode: Nulls

The PostalCode contains close to 50% NULL values. This is significant because this accounts for 25,000+ NULLs within the column.

Table

Description automatically generated

Postal code is an important piece of the complete address. Street number and name for a business is sufficient if street names are completely unique but that is not the case within the city or the country. There are many duplicates therefore a postal code is necessary.

Business names are also not always unique or a business may have multiple locations. Therefore a complete address is necessary.

Additionally, if anyone wants to get in contact with the business though mail, they would not be able to get in contact due to lack of postal code.

A screenshot of a computer

Description automatically generated with medium confidence

This image was generate through a pivot table. It shows a breakdown for the Null values for PostalCode. The vast majority of business with null postal codes have been issued their business license. This shows that there is incomplete proof of address recorded for businesses that were still running.

A screenshot of a computer

Description automatically generated

Out of the businesses that were issued their licenses, a vast majority of the null values for postal code are from the housing sector (strata, detached house, secondary suite). Contractor and office space are also significant portion of this issue.

## PostalCode: Format

As highlighted earlier, a postal code is necessary for a complete address. All Postal codes in Canada follow the convention LDL DLD (L = letter, D = digit). All records in the PostalCode column do not follow this format.

It would also be helpful to determine if these postal codes are only for Canada or if they could also include equivalent values from other counties. Currently I am assuming that these are only Canadian postal codes.

A picture containing text, screenshot, number, font

Description automatically generated

Ideally a postal code should have a space between LDL and DLD as they are both meaningful metrics which make up the postal code. About 300 records do not have correct spacing between the two segments. Additionally, many postal codes do not follow the standard convention and are therefore invalid entries. Some do not have the proper sequence of LDL DLD, others contain symbols, and some are missing characters.

All these issues indicates that is an underlying issue that corresponds to the correct format of the postal code for a business.

A check should exist when inputting the postal code as the convention is standard. This check would ensure correct number of characters, format for letters and digits, and proper spacing.

## LicenceRSN: Duplicate

LicenceRSN, as mentioned on Open Data Vancouver, is the unique primary key. This means that each RSN should be unique causing no duplicates.

A screenshot of a graph

Description automatically generated with low confidence

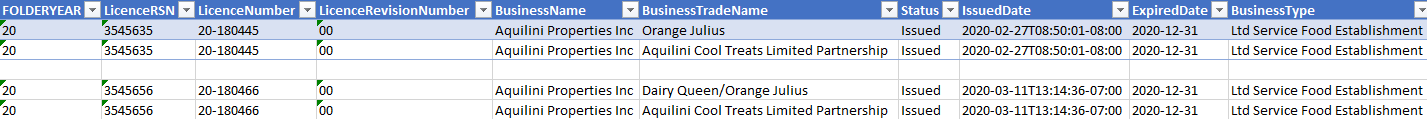
Two RSN values are duplicates in the database.

When trying to change or update the business’ status, it is likely that LicenceRSN will be used as it should not contain duplicates. Changing the records through this RSN affect each record with that particular RSN. This means that a row will be mistakenly changed as it shares the key with another.

A screenshot of a computer

Description automatically generated with low confidence

This issue exists for the same business twice.



At both times Aquilini Properties Inc obtained licenses for Aquilini Cool Treats and Orange Julius. There may be an error occurring when obtaining two licenses at once unless this is the correct response for when a business operates multiple different operations under its name at the same property.

## Province: Format

The province column stores the province portion of an address. There is no particular format that this column uses which is resulting in provinces either being listed more than one in different forms, or inputs not being province. This is an issue because the of a lack of convention.

Additionally it is not certain that only provinces from Canada can be displayed here as the column name indicates that it may only be province and not equivalents of other countries

Provinces in Canada can follow a 2 letter format such as LL (L = letter). It is not entirely compulsory to have a 2 letter format unlike the postal code, but the convention helps to ensure that the correct province is inputted.

A screenshot of a table

Description automatically generated with low confidence

AB and Ab inputs show that one province is shown two different ways due to capitalization differences. British Columbia and BC also appears in two formats. There exist some inputs such as NY, WA which may be referring to states in the USA. There also exist other inputs such as CA and IN which may refer to states or country codes.

Additionally there exist records that are not a province or a state such as Vancouver and Lakeview Estates.

Due to many formatting issues, it would be difficult to aggregate on this column as it would provide with inaccurate numbers.

## Status: Invalid input

This value is assigned in order to keep track of the stage of processing the license as well as status of the business.

The status can be one of:

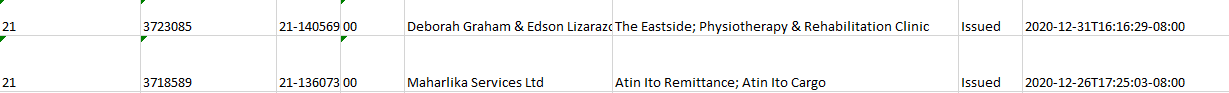
* Issued – business license is issued for the year relating to the issue date
* Pending – this can either mean that a license is incomplete or in review
* Inactive – License is not active
* Cancelled - license is cancelled (no reason is provided in the dataset)
* GOB – the business has gone out of business

99.99% of the 51,000+ rows have the correct status associated with the business license.

A screenshot of a computer

Description automatically generated with low confidence

There exist 2 anomalies within the column which seem to be business names. The two business names only occur once therefore it is likely not an issue that results repeatedly for the same business. 



Both of these error does not seem to be related to issuing the license on the same day because they differ in days, although they share the month so it could be a factor. Both the businesses display fine on Excel with Status containing the correct input. The similarity they share is the semicolon in BusinessTradeName. It is possible that the use of the semicolon resulted in the analyzer mistake the correct format for those two rows.

# **SME Review of Initial Assessment**

## PostalCode: Null

25,000+ Null values for a column is significant yet it is a problem that would most likely not be feasible for the current year. A fix should be in works for next year’s cycle on issuing business licenses and ensuring postal code is filled out.

The only way to fill out the null values would be to contact every business to gather their accurate information which is wasteful of resources. Therefore the priority would be **MEDIUM** because we need a solution for next year but should not be chasing to complete the current database.

## PostalCode: Format

By the looks of it there may not be checks in place to ensure that correct format for the postal code is followed. Out of the 25,000 postal codes that were filled, only about 50 do not follow the correct convention, therefore resulting in an incorrect postal code.

The priority of this should be **LOW** because this issue persists for less than 0.2% of the records and checks could be set in place before next year.

## LicenseRSN: Duplicate

The priority is set to **LOW** as this has only occurred on a rare occasion and RSN has been unique for all other cases. This occurred for the same business on two occasions that runs multiple businesses.

It would be beneficial to see if there was an influx of licenses being issued that day as this could indicate if higher batch could form an exploit in assigning RSN. A time series line graph would aid in seeing this.

## Province: Format

The priority is set to **LOW** as these issues are only present in less than 0.04% of the records. A check can be put in place to select province from a dropdown menu and an additional space to input if business is located outside of Canada. This could include a dropdown menu to choose for the provinces resulting in a better convention used. It can also include another option that allows for a inputting a string for addresses outside of Canada.

## Status: Invalid input

The priority is set to **LOW** as this seems like an anomaly. This issue is not repeated for any other row apart from the 2 out of the 50,000. It would be good to test out the use of semicolon to determine if it causes an issue in analyzing the data by using dummy rows.

# **Further Research for SME**

## LicenseRSN: Duplicate

This portion aims to look deeper at LicenseRSN consisting of duplicate values. This had occurred twice for Aquilini Properties Inc. It is beneficial to see the count of RSN for that day to determine if increase in issuing RSN resulted in this error.

A picture containing text, screenshot, line, plot

Description automatically generated

The count for licenses being issued for the two days, 2020-02-27 & 2020-03-11, that resulted in duplicate RSN were not particularly high. Rather those days were lower than the average licenses issued. This can be shown by the green rectangle in the graph

The average issue count for the year was 144, meanwhile the count on the two days were 108 and 73 respectively. The issue for duplicates doesn’t seem to be caused by high traffic on those particular days.

A screenshot of a spreadsheet

Description automatically generated with medium confidence

This table shows that for the month of February, the count was not significantly greater than the average, and the month of March was less than the average. Both the daily and monthly findings indicate that it is unlikely that duplicate RSN was the result of too many licenses being issued, rather it may be something else.

SPC for Service Food Establishment business type, the business type of Aquilini Properties Inc also shows that licenses being issued for that sector was below average for the two days that duplicates were issued. The average for the year was 7, meanwhile the count was 6 and 3 for the respective days.

# **SME Review of Further Research**

The average count was lower than the average so it is likely that this was not due to system being overloaded. The numbers look under statistical control for the dates concerned. Since the duplicate RSN was assigned for the same business twice, there may be other policies at stake that resulted in this which should be investigated.

The graph does present another challenge. It seems that businesses are more likely to obtain license towards the end of the year. This can result in an overload, particularly the time the count overshoots the upper ceiling by quite a bit.

We can look at ways to incentivize businesses to obtain their license quicker to help with the increase at the end of the year.

# **SME Suggestions for DQ Rules**

## PostalCode: Format and Null

A check should be put in place which ensures that this field is not left empty if the status of the record is “Issued”as the postal code is an important part of the address as street names are not unique. This would ensure that it is not left null.

Another check should be in place that will make sure that the format of the postal code follows the national convention for Canada. This would be “LDL DLD” (L = letter, D = digit). A person could still put in the wrong postal code due to an error on their part but it now makes it easier to have the right one inputted. For companies that may not be based in Canada, the equilvant of a postal code (pincode, zipcode) should follow convention of that particular country.

Since most businesses are in Canada an option should be present where if country is not Canada then it allows a person to have more freedom when writing the address as conventions would not be the same.

## LicenseRSN: Duplicate

We need to revisit policy and check if it can allow for duplicate RSN. If that is the case then it should be revised as duplicates in a database are an issue, especially for RSN as it is supposed to be unique in every instance. This can be resolved by automating the process of assigning RSN and having it be unique each time. If RSN can be edited, then a check should exist to show that a duplicate already exists and therefore the edit is unsuccessful and another number should be picked.

For either case of automating or editing, an alert should be signaled if the LicenseRSN matches another

## Province: Format

A dropdown menu to pick provinces/territories should be present with the standard 2 letter convention. The amount of options are limited and static so unique inputs should not need to exist. If a company is not from Canada then two capitalized rule for province should not be in place and input of string specifying state/province be allowed.

## Status: Invalid input

The status should match one of the options and not allow input of another string. The expiration date should also be checked against current date and if current date is greater than expiration date then status should change if it is currently set as “Issued” to “Inactive”