

**Specialist Diploma in Business Analytics**

**Specialist Diploma in Financial Analytics**

**AY2020/2021 Apr Semester**

**Business Intelligence Fundamentals**

**(CBA1C07)**

**Assignment**

Submitted by

Ms Wee Kar Ghee(2080985A)

Date of Submission : 15 June 2020

# 

# 

**Specialist Diploma in Business Analytics**

**Specialist Diploma in Financial Analytics**

**Business Intelligences Fundamentals (CBA1C07)**

**AY2020/2021 Apr Semester**

**Assignment**

**Submitted by:** 2080985A | Wee Kar Ghee

**Date:** 15/06/2020

“By submitting this work, I am declaring that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person’s work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism.”



Name and Signature of student: …………………………………...

Ms Wee Kar Ghee

**Table of Contents**

1 PROJECT OVERVIEW ........................................................................................................................ 1

2 METHODOLOGIES ............................................................................................................................ 1

3 DATA PROFILING ............................................................................................................................. 1

3.1 DATA QUALITY ................................................................................................................. 1

3.2 PROFILING OF ATTRIBUTES .............................................................................................. 2

3.3 SUMMARY ........................................................................................................................ 3

4 DATA CLEANING ............................................................................................................................... 3

4.1 ADDITION OF ATTRIBUTES……………………………………………………………………………………………. 3

4.2 MISLABELLED CLASSES....................................................................................................... 4

4.3 MISSING VALUES & OUTLIERS ........................................................................................... 5

4.4 ATTRIBUTE – ACTION NOT TAKEN AT THIS STAGE............................................................. 5

5 DATA INTEGRATION ........................................................................................................................... 6

6 DATA EXPLORATION ........................................................................................................................... 6

7 SUMMARY ......................................................................................................................................... 10

Annex A – Legend for FRANCHISE\_500 and FRANCHISE\_CEO .............................................................. 11

Annex B – Data Profiling on all attributes on FRANCHISE\_500 and FRANCHISE\_CEO........................... 13

Annex C – Screenshots of formulaes used for calculated columns....................................................... 13

Annex D – Histogram of Change In Units 1 Yr, Change In Units 3 Yr, Net Worth Requirement, Liquid Cash Requirement ………………………………………………………………………………...................................................... 15

1 PROJECT OVERVIEW

Businesses had been hit badly by the pandemic Covid-19. It has been a race for modifications and survival tactics. What type of business can survive in this aftermath? What type of businesses were doing well previously? How much was the gain? Moving forward, how to get stakeholders to invest in franchise business?

There was a study on the franchise industry. Two datasets were collected through gathering of the most comprehensive franchise ranking for franchisees in the U.S. or Canada only. However, with each year, the ranking becomes increasingly competitive, as more and more companies see the value of expansion through franchising and throw their hats into the ring. The list shows just how varied the franchise world has become, offering prospective franchisees a myriad of choices, from more traditional options, like restaurants, maintenance services, and gyms to the latest trends. Hence, evaluating and ranking from a diverse pool of opportunities has becoming more challenging.

As a business consultancy firm, our team was tasked to find out some insights on the trends of franchises.

2 METHODOLOGIES

The two datasets to be used for this case study are FRANCHISE\_500 and FRANCHISE\_CEO. For analysis purpose, the Power BI software will be used.

In order to quantify the increase or decrease of number of franchise units, some methodologies to be used to estimate the change in franchise units in one or three years and the net worth requirement and liquid cash requirement.

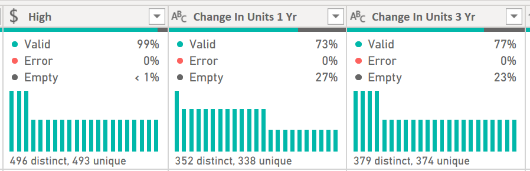
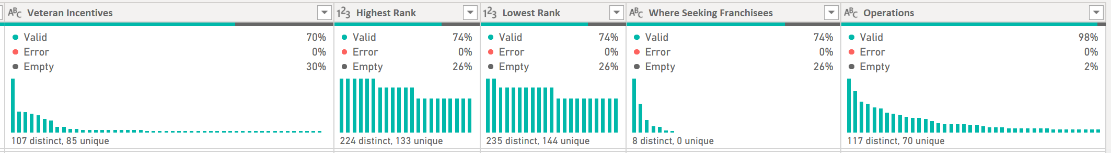
* A base of 1,000 is used for change in units 1 Yr. Therefore, if Change in Units in 1 Yr is +0.7% + 11 units, change in franchise units in 1 Yr = 0.7/100 1,000 + 11 = 18 franchise units. For null values in change in franchise units it can be replaced with overall average value.
* A base of 1,000 is used for change in units 3 Yr. Therefore, if Change in Units in 3 Yr is +9.6% + 7 units, change in franchise units in 3 Yr = 9.6/100 1,000 + 7 = 103 franchise units. For null values in change in franchise units it can be replaced with overall average value.
* The average is used for given range of net worth requirement and liquid cash requirement. Therefore, if net worth requirement is $80,000 - $310,000, the net worth requirement = (80,000 + 310,000)/2 = $195,000. For null net worth and liquid cash requirement, it can be replaced with the overall average value.

3 DATA PROFILING

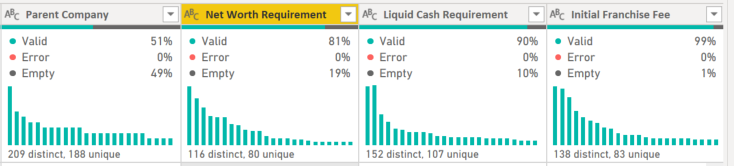
In this section, we will be discussing the data quality and profiling of the two datasets.

3.1 DATA QUALITY

*Figure 1: FRANCHISE\_500 - Screenshots of “Column Quality” results in Power BI for the 8 attributes with dirty data.*

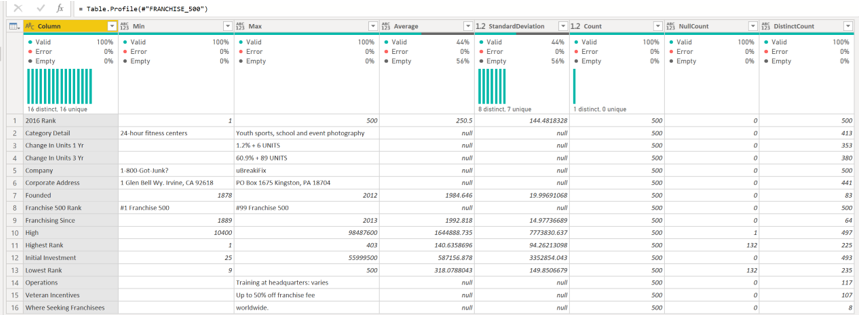


*Figure 2: FRANCHISE\_CEO - Screenshots of “Column Quality” results in Power BI for the 4 attributes with dirty data*

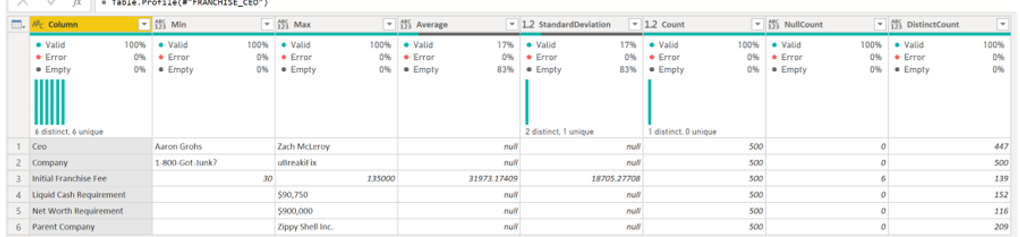


While examining the data quality using Power BI, it is noticed that:

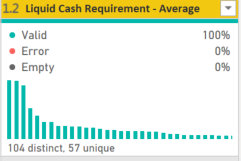
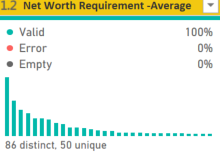
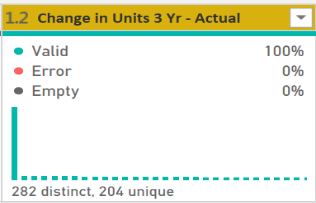
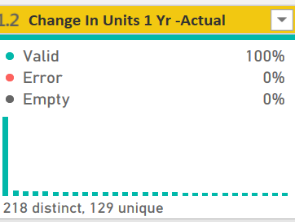
* In *FRANCHISE\_500*, there are 100% valid data values while 8 attributes contained dirty data as shown in Figure 1. The 8 attributes have missing values that amounted to ranging from less than 1% to approximately 30% of the data.
* In *FRANCHISE\_CEO*, there are 100% valid data values while 4 attributes contained dirty data as shown in Figure 2. The 4 attributes have missing values that amounted to ranging from approximately 1% to 49% of the data.
  1. PROFILING OF ATTRIBUTES



*Figure 3: Screenshot of “Summary Statistics” for FRANCHISE\_500 in Power BI*

*Figure 4: Screenshot of “Summary Statistics” for FRANCHISE\_CEO in Power BI*

As shown in Figure 3 and 4, in the initial representation of the Summary Statistics for FRANCHISE\_500 and FRANCHISE\_CEO, it does not show any anomalies. However, it is noted that for ‘Change in Units 1 Yr’, ‘Change in Units 3 Yr’, Net Worth Requirement and Liquid Cash Requirement, we need to use the methodologies to obtain the absolute value. The steps of getting the absolute value will be explained in data cleaning section.



|  |  |
| --- | --- |
| Count | 500 |
| Error | 0 |
| Empty | 0 |
| Distinct | 218 |
| Unique | 129 |
| NaN | 0 |
| Zero | 0 |
| Min | 4 |
| Max | 2970 |
| Average | 160.014 |
| Standard deviation | 211.171372351941 |

|  |  |
| --- | --- |
| Count | 500 |
| Error | 0 |
| Empty | 0 |
| Distinct | 282 |
| Unique | 204 |
| NaN | 0 |
| Zero | 0 |
| Min | 2 |
| Max | 5592 |
| Average | 291.186 |
| Standard deviation | 374.388152030431 |

|  |  |
| --- | --- |
| Count | 500 |
| Error | 0 |
| Empty | 0 |
| Distinct | 104 |
| Unique | 57 |
| NaN | 0 |
| Zero | 0 |
| Min | 1000 |
| Max | 2000000 |
| Average | 162917.125 |
| Standard deviation | 223371.281175956 |

|  |  |
| --- | --- |
| Count | 500 |
| Error | 0 |
| Empty | 0 |
| Distinct | 86 |
| Unique | 50 |
| NaN | 0 |
| Zero | 0 |
| Min | 500 |
| Max | 50500000 |
| Average | 534596.051 |
| Standard deviation | 2293167.49808328 |

*Figure 5: Screenshots of the absolute value of Change In Units 1 Yr and Change In Units 3 Yr*

*Figure 6 : Screenshots of the Net Worth*

*Requirement and Liquid Cash Requirement*

After obtaining the absolute value, as seen in Figure 5 and 6, it was noted that the maximum value is more than three standard deviations away from the mean (average value). Hence, it is a common practice to identify them as outliers.

* 1. SUMMARY

In summary, the two datasets contained dirty data and the information is as tabulated in Table 1. The full list of observations for all the attributes are included in Annex B. To note, the ‘2016 Rank’ attribute in FRANCHISE\_500 represented the rank of each companies and it is also the unique ID. The unique value is 500 which tallies with the unique values of ‘Company’ attribute in the two datasets, it is concluded that there are no duplicate of records.

*Table 1 : Summary table on the dirty data in the two datasets.*

|  |  |
| --- | --- |
| **Attribute Name** | **Observations** |
| Category Details (FRANCHISE\_500) | Many distinct values. |
| Franchise 500 Rank (FRANCHISE\_500) | Same information with attribute ‘2016 Rank’. |
| High | Missing value. |
| Change In Units 1 Yr(FRANCHISE\_500) | Formulae is used, no absolute value. One value is different and does not have ‘+’ sign in front. |
| Change In Units 3 Yr(FRANCHISE\_500) |
| Veteran Incentives (FRANCHISE\_500) | Missing values. |
| Highest Rank (FRANCHISE\_500) | Missing values. |
| Lowest Rank (FRANCHISE\_500) | Missing values. |
| Where Seeking Franchisees (FRANCHISE\_500) | Mislabelled classes. |
| Operations (FRANCHISE\_500) | Many distinct values. |
| Parent Company (FRANCHISE\_CEO) | Missing values. |
| Net Worth Requirement (FRANCHISE\_CEO) | Inconsistent format as some are single values and some values are shown in a range. Also, there are missing values. |
| Liquid Cash Requirement (FRANCHISE\_CEO) |
| Initial Franchise Fee (FRANCHISE\_CEO) | Missing values. |

1. DATA CLEANING

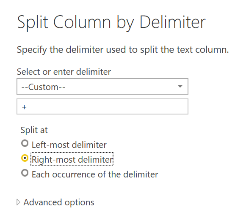
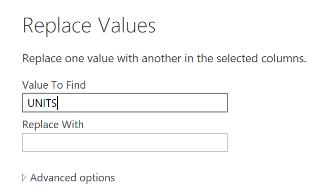
In this section, the data problems will be resolved using the data cleaning tools in Power BI.

4.1 ADDITION OF ATTRIBUTES

It was observed that in FRANCHISE\_500 and FRANCHISE\_CEO, there was a need to find out the absolute value in the growth of the businesses. To facilitate further data analysis, new attributes are added to the data and they are tabulated in Table 2.

*Table 2: Explanations for new attributes to be added.*

|  |  |
| --- | --- |
| **Attribute Name** | **Explanation** |
| **FRANCHISE\_500** | |
| Change In Units 1 Yr - % | * These two additional columns were created for calculation to get absolute value of Change In Units 1 Yr   *(Figure 7 & 8 on Replace and Split Column in Power BI)* |
| Change In Units 1 Yr -UNITS |
| Change In Units 1 Yr - Actual | * This attribute will show the absolute value for Change In Units 1 Yr. * ‘= [Change In Units 1 Yr -%]\* 1000 + [Change In Units 1 Yr – UNITS] ‘. *(Annex C for screenshot)* * Overall Average was calculated as ‘160’.      * Null values were replaced by ‘160’. |
| Change In Units 3 Yr - % | * These two additional columns were created for calculation to get absolute value of Change In Units 3 Yr. |
| Change In Units 3 Yr -UNITS |
| Change In Units 3 Yr - Actual | * This attribute will show the absolute value for Change In Units 3 Yr. * ‘= [Change In Units 3 Yr -%]\* 1000 + [Change In Units 3 Yr – UNITS] ‘. *(Annex C for screenshot)* * Overall Average was calculated as ‘291’.      * Null values were replaced by ‘291.’. |
| Franchising Since by Decade | * It will be good to know see trends by decade, eg. 1980s, 1990s, etc. * = Text.Start([Franchising Since],3) & “0s” *(Annex C for screenshot)* |
| Operations by Location | * To reduce distinct values and to view the training locations. * Split from attribute, Operations. * 4 distinct values are left after the split. * Missing fields replaced with ‘NA’. |
| Operations By Frequency | * Split from attribute, Operations. * Distinct values are still many as there is no standardized format (eg. 2 weeks+, 135 hours, varies). Hence, we will need to check with data owner on the precise or valid values. This will not be the priority at this stage. * Missing fields replaced with ‘NA’. |
| **FRANCHISE\_CEO** | |
| Net Worth Requirement- Min | * Some of the values are in range format. For such format, average value will be used. * These two attributes were split from attribute, Net Worth Requirement. |
| Net Worth Requirement -Max |
| Net Worth Requirement -Average | * Select attributes Net Worth Requirement- Min and Max. Add Column ->  -> from dropdown list, select Average. * We need to find the overall Average and replaced the null values. * Click on ‘Net Worth Requirement -Average’. Transform -> Statistics -> Average.      * Replace null values with overall Average **‘534596’.** |
| Liquid Cash Requirement - Min | * Some of the values are in range format. For such format, average value will be used. * These two attributes were split from attribute, Liquid Cash Requirement. |
| Liquid Cash Requirement - Max |
| Liquid Cash Requirement – Average | * Same steps as ‘Net Worth Requirement -Average’. * Overall Average is 162917.      * Replace null values with overall Average **‘162917’.** |



*Figure 7: Screenshot of split Column by Delimiter Figure 8: Screenshot of Replace Values*

* 1. MISLABELLED CLASSES

|  |  |
| --- | --- |
| **Attribute Name** | **Explanation** |
| Where Seeking Franchisees | * Supposed to have 3 distinct values but there were 7 which signified mislabelled classes. This will be rectified via calculated formulae   *(Annex C for screenshot)*.   * The missing values will also be replaced by ‘unknown’. |

* 1. MISSING VALUES AND OUTLIER

There are quite a number of missing values and they appeared in different rows. Deletion of these missing values may caused distortion and misinterpretation of data. Hence, the missing values will be treated as follows:

|  |  |
| --- | --- |
| **Attribute Name** | **Explanation** |
| High | Replaced 1 null field with ‘0’. |
| Veteran Incentives | Replaced 152 missing values with ‘NA’. |
| Highest Rank | Replaced 132 null fields with ‘0’. |
| Lowest Rank | Replaced 132 null fields with ‘0’. |

An outlier is a data point that is more than three standard deviations away from the mean value and for this case study, all outliers will be removed. To determine the values of the outliers for Change In Units 1 Yr, Change In Units 3 Yr, Net Worth Requirement and Liquid Cash Requirement. Histograms were plotted. The screenshots of the histograms are included in Annex D.

* 1. ATTRIBUTE – ACTION NOT TAKEN AT THIS STAGE

|  |  |
| --- | --- |
| **Attribute Name** | **Explanation** |
| **Category Detail** | There are many business trends. It will be good to re-categorize broadly. For example, salads, burgers, restaurants to be grouped under ‘Food’. However, it is noted that this is not the priority at this stage. In addition, we will need to check with data owners on criteria for groupings. |

1. DATA INTEGRATION

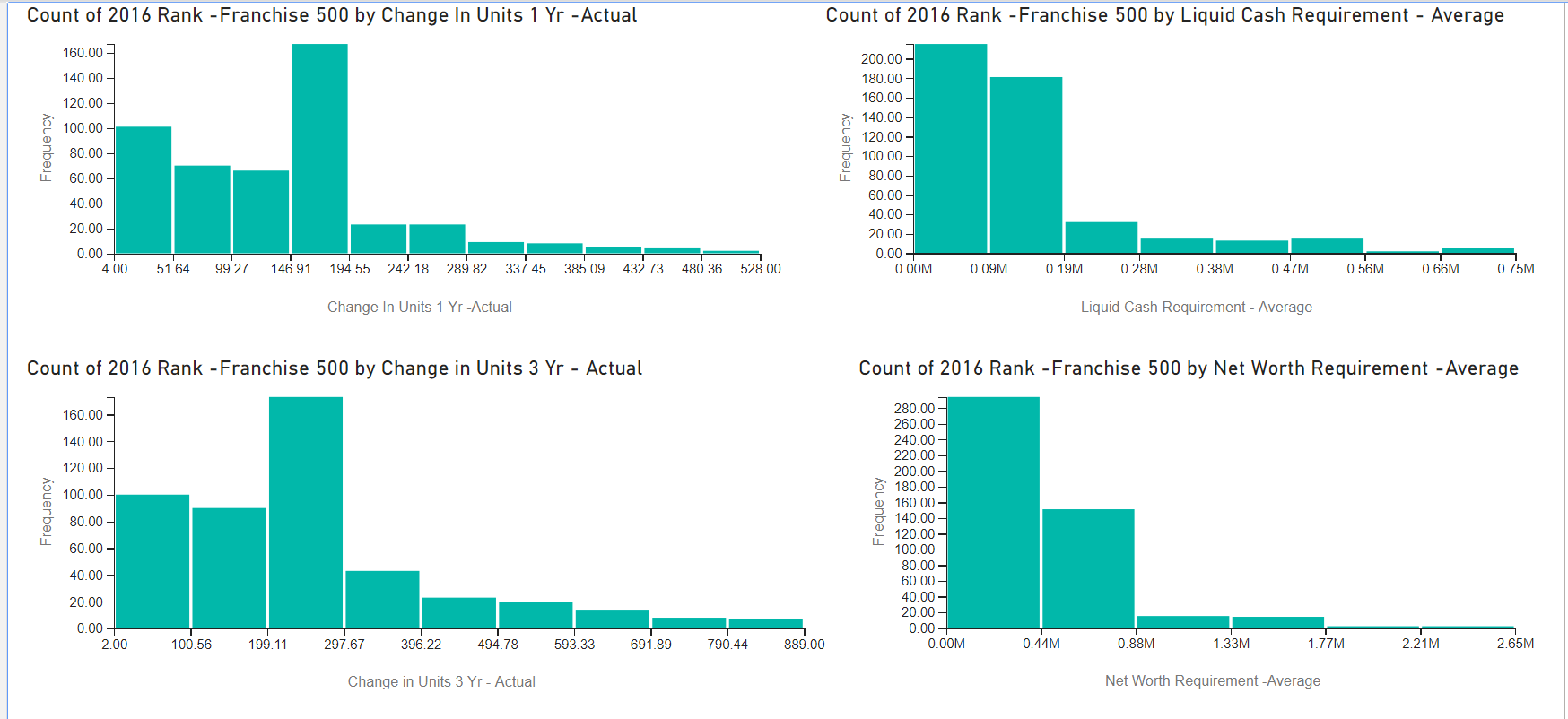
After cleaning the two datasets, the next step was to link the two datasets together as shown in Figure 9. The join key is ‘Company’.

Figure 10, we will then proceed to merge the two datasets prior to data exploration. FRANCHISE\_CEO to merge with FRANCHISE\_500.

|  |  |
| --- | --- |
| *Figure 9: Link the two tables in Power BI* | *Figure 10: Merge the two datasets in Power BI* |

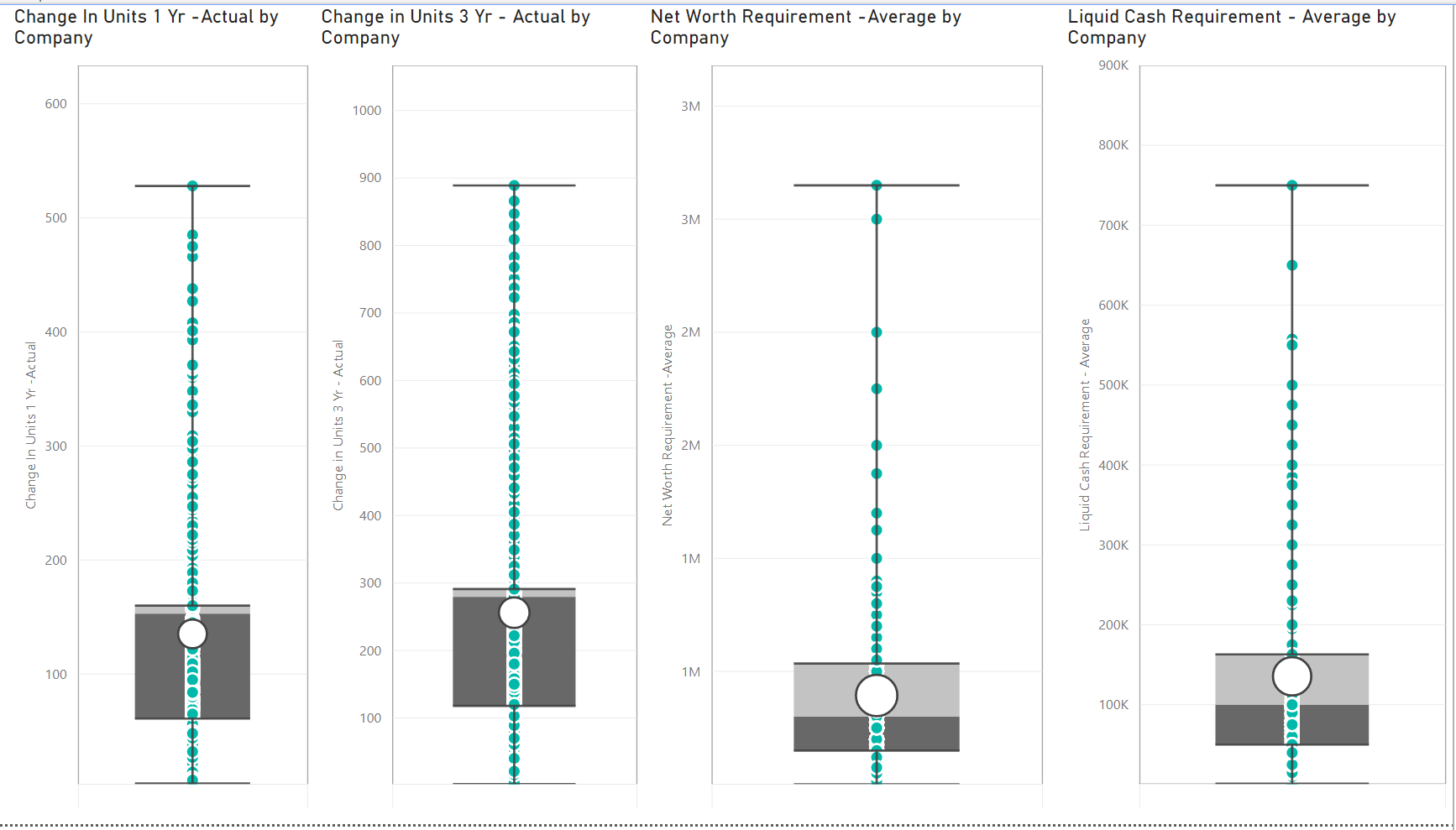
1. DATA EXPLORATION

*Figure 11: Histogram of Change In Units 1 Yr, Change in Units 3 Yr, Liquid Cash Requirement and Net Worth Requirement*



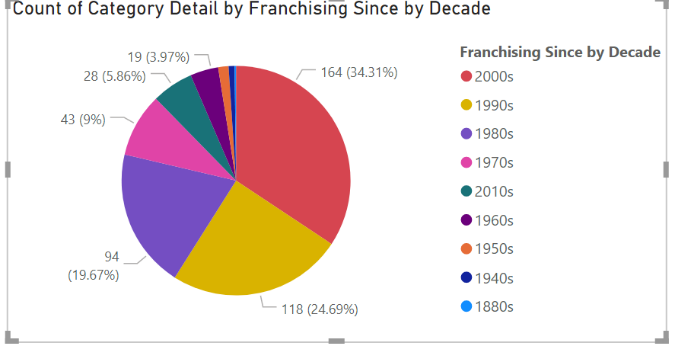
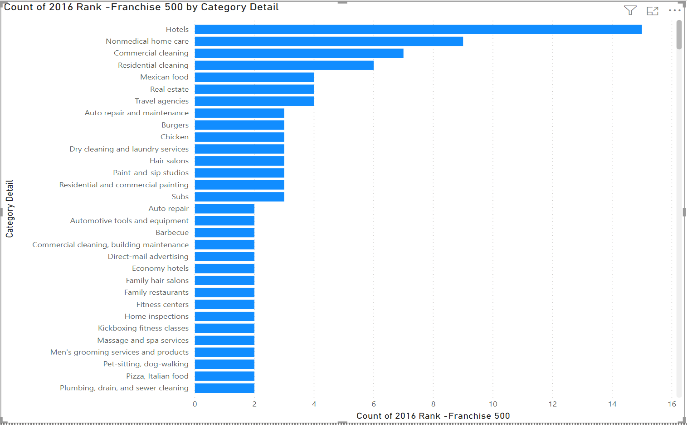
As shown in figure 11, most of the franchisees have growth of not more than 194 units in 1 Yr and less than 297 in 3 Yr. The non-borrowed and cash on hand liquid cash requirement from franchisees were mostly 0.19million or less. The net worth requirement from franchisees was mostly at 0.88million or less. Data distribution for all is right skewed or positively skewed with spread more to the right which also indicated that the mean is more than the median.

*Figure 12: Box Plot of Change In Units 1 Yr, Change in Units 3 Yr, Liquid Cash Requirement and Net Worth Requirement*



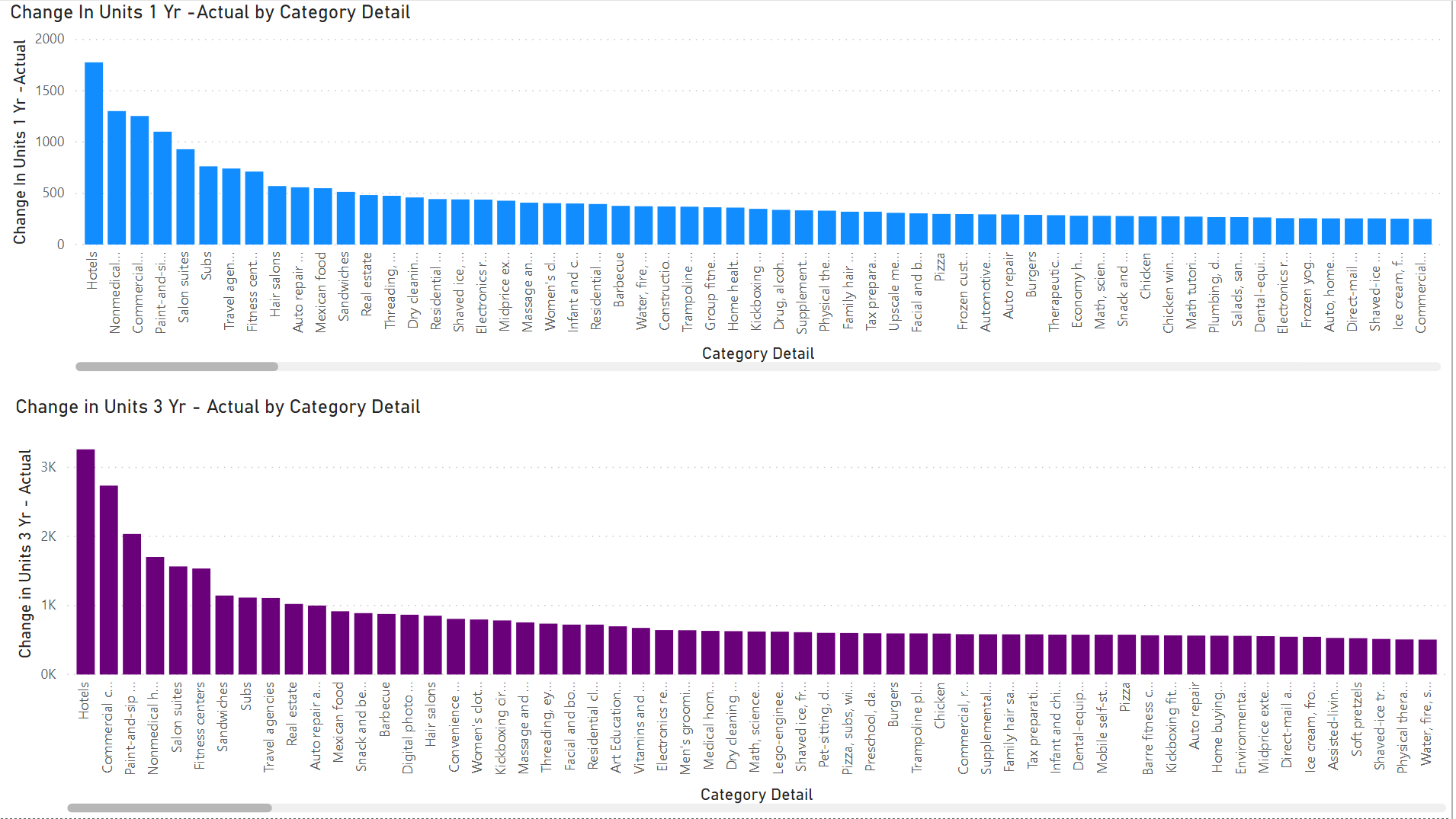
As seen in figure 12, sections of box are uneven in ‘Change In Units 1 Yr and Change in Units 3 Yr’ which can be interpreted that the below median are spread widely while above median are bunched together. It is also noted that the franchisees’ Liquid Cash Requirement is a little more consistent as compared to the Net Worth Requirement.

*Figure 14: count of Category Detail*

*Figure 13: Category Detail by Decade*

It is interesting to see from figure 13, the categories had grown diversely since 1880s with 164 categories in 2000s, 28 categories in 2010s and 1 category (Language, Intercultural and business Training) in 1880s. As seen from figure 14, Hotels has the highest number of franchisees.

*Figure 15: Change In Units 1 Yr and 3 Yr by Category Detail*



Hotel has the highest number of franchisees and thus has more units growth in 1 Yr and 3 Yr. With more hotels, it probably resulted in more demand for cleaning services which can be seen from figure 15, Nonmedical home care was ranked in no.2 in 1 Yr but dropped to no.4 in 3 Yr. Commercial Cleaning had grown with more units in 3 Yr taking rank no.2.

*Figure 16: Change In Units 1 Yr and 3 Yr by Company*

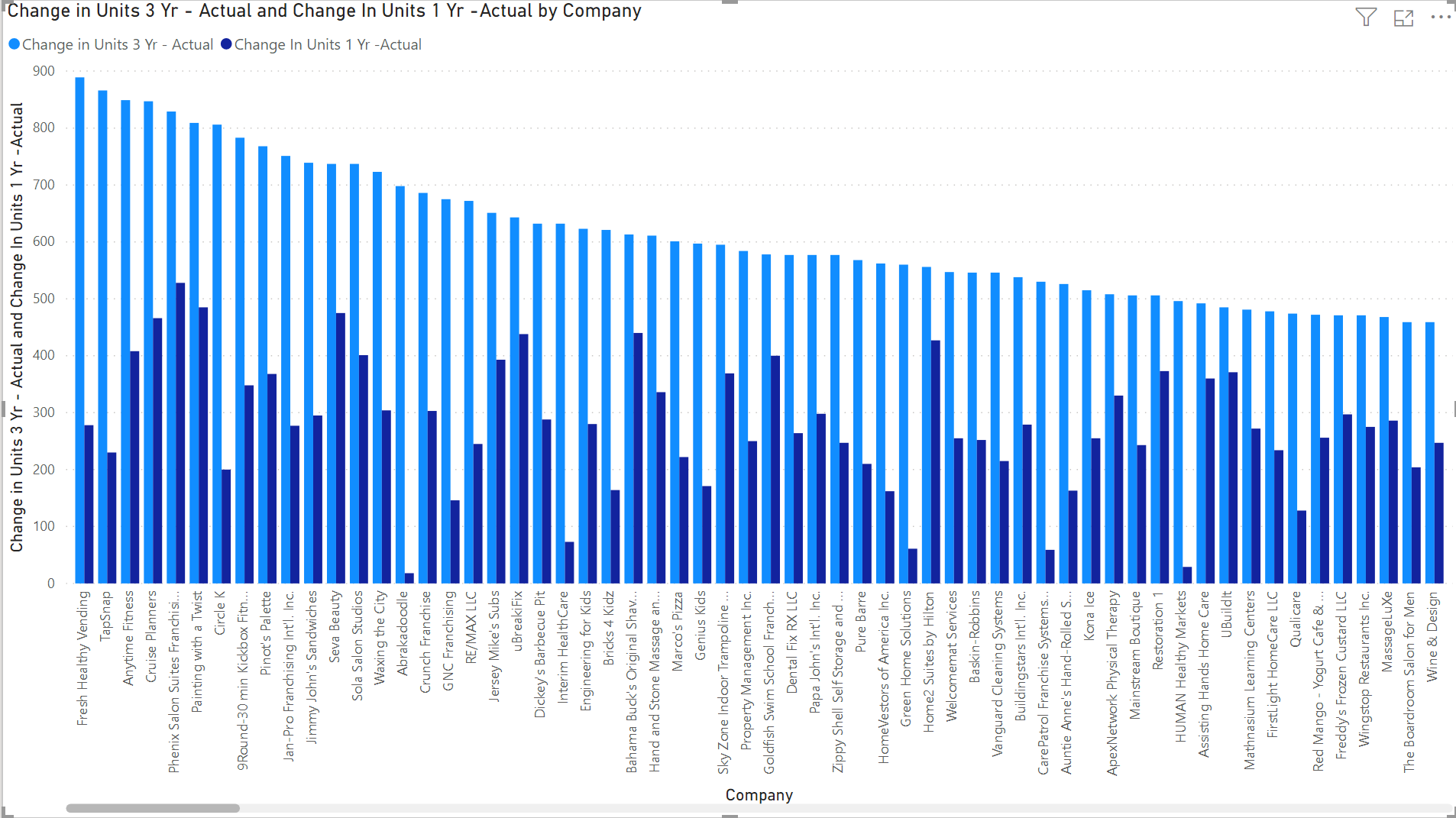


Figure 16, Hotel may have the highest number of franchisees but in view of increase in Units by Company from 1 Yr to 3 Yr, F&B, beauty and fitness companies took the top 3 places.

*Figure 17: Locations where Franchising in available*

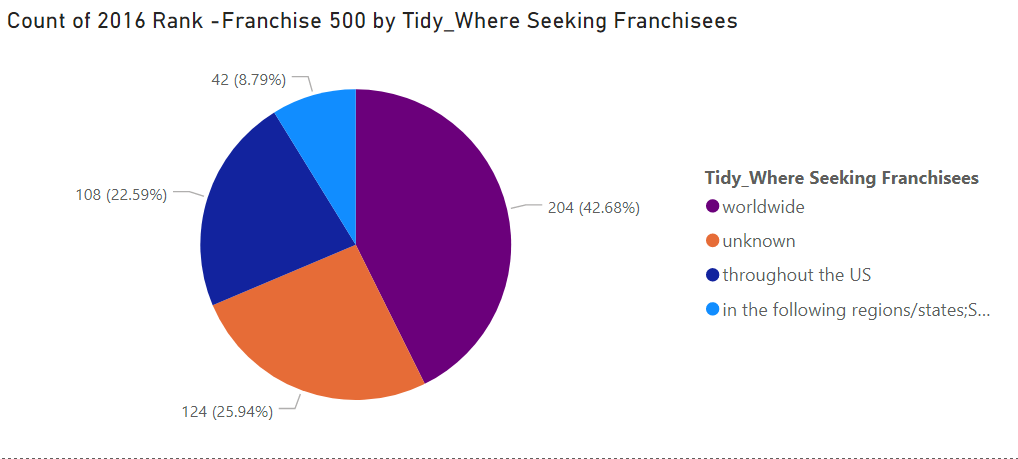


Figure 17 shows that there are a lot of opportunites for franchise business in worldwide.

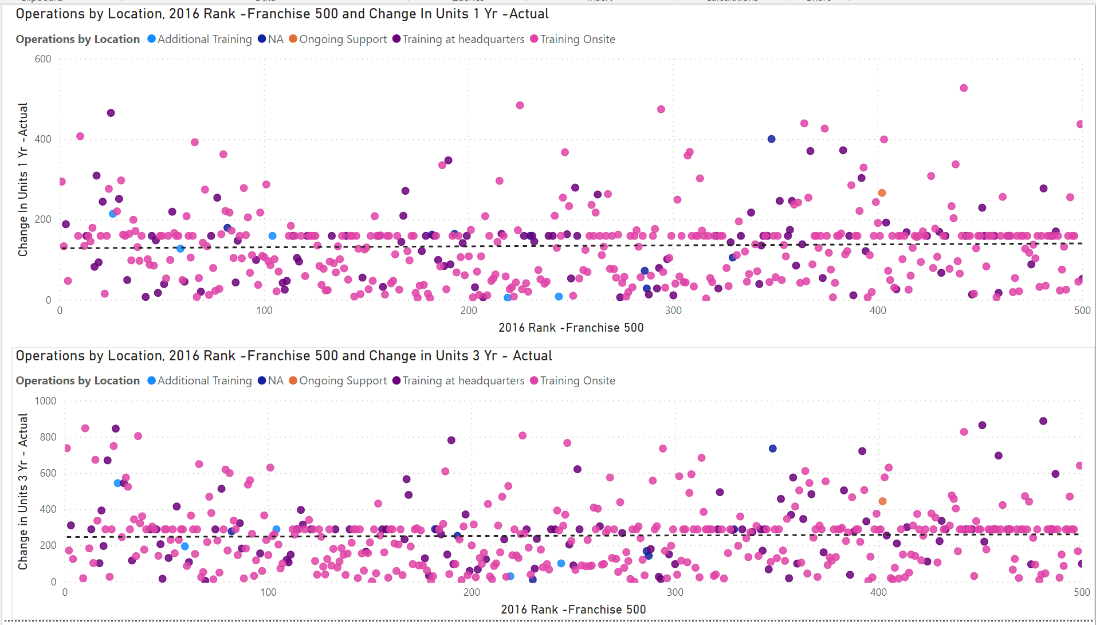
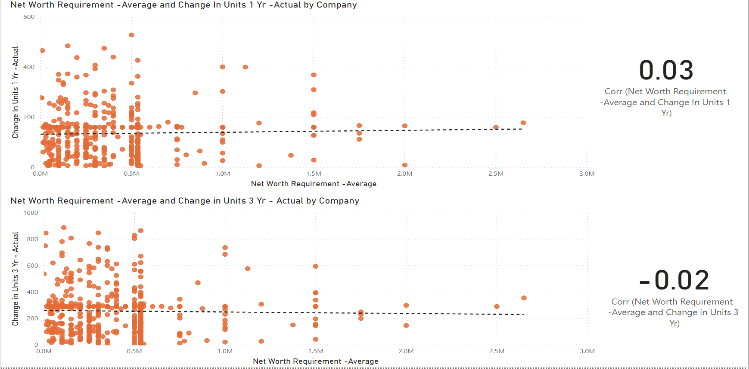
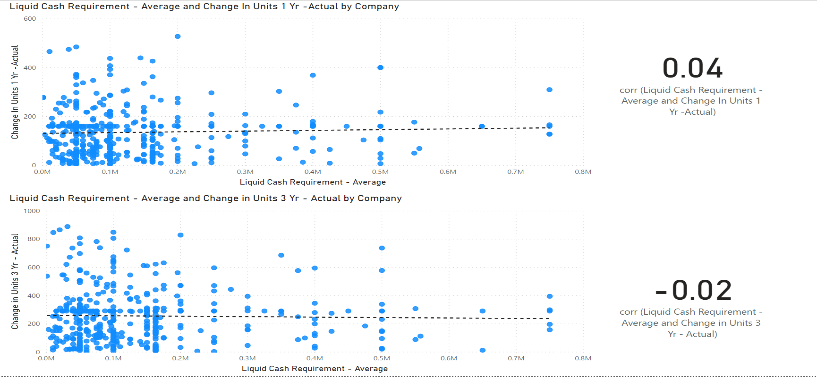
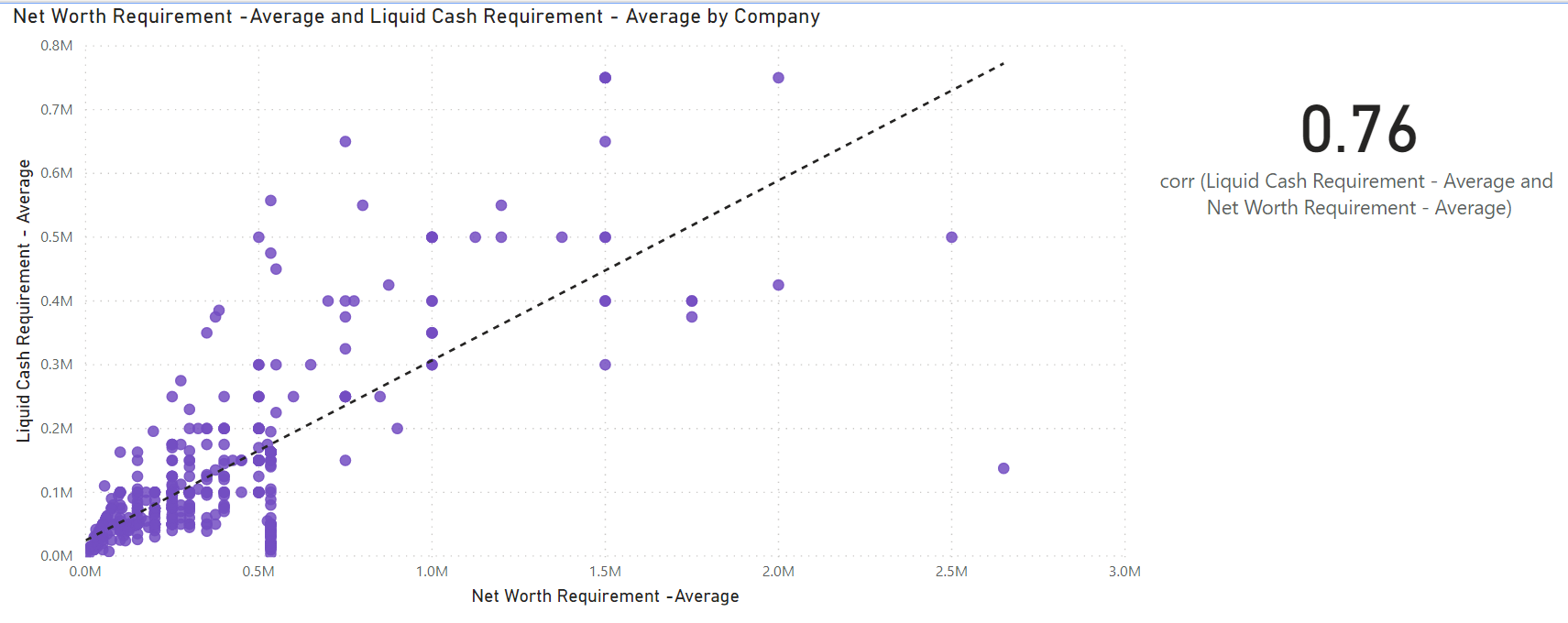
*Figure 18: Training location vs Change In units*

Figure 18, there is no relationship between training location and growth in Units for 1 Yr and 3 Yr.



*Figure 19: Net Worth Requirement vs Change In Units Figure 20: Liquid Cash Requirement vs Change In Units*

The two mandatory requirements for a person to have in order to qualify to become a franchisee are Net Worth Requirement (minimum net worth) and Liquid Cash Requirement (minimum available liquid capital). Figure 19 and 20, these two requirements have almost no relationship with growth of business which is Change in units.

*Figure 21: Net Worth Requirement vs Liquid Cash Requirement*

The Net Worth Requirement and Liquid Cash Requirement have a direct and linear relationship with moderately strong and positive strength of 0.76.

7 summary

The findings of FRANCHISE\_500 was not intended to endorse, advertise or recommend any particular franchise. The intent tallies with the findings from this case study. It was observed that Hotel category may be the most popular trend among franchisees, but was not the top 3 companies in growth in Units. Instead, F&B, beauty and fitness companies took the top 3 places in Change In Units 1 Yr and 3 Yr. The category detail could also be further re-grouped for more indepth analysis but will need to check with data owners for criteria.

For one to start franchising business, it is noted from the findings that the two mandatory requirements; Net Worth Requirement and Liquid Cash Requirement have no impact on business growth. Training location also does not have relationship with business growth but we can’t conclude on the data for frequency of training as the data needs to be further clarify with data owners for re-grouping.

To conclude, there are a lot of opportunities for franchising in worldwide. The above findings will be a guideline to present to stakeholders. However, ulitmate decision will still lie with them. More research may need to be done before investing in a franchise.

Video link: <https://youtu.be/m6QE_CjwCsY>

**Annex A – Legend for FRANCHISE\_500 and FRANCHISE\_CEO**

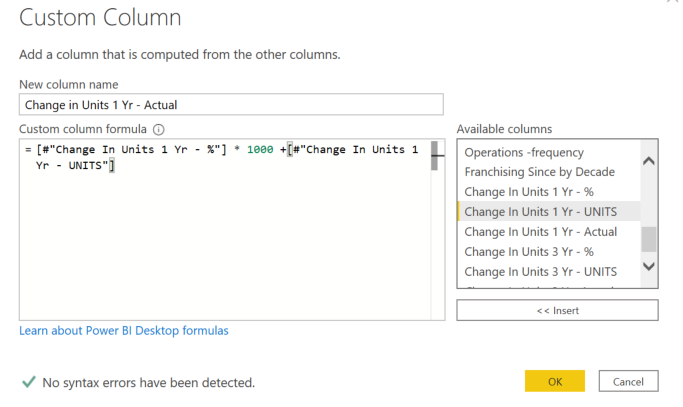
|  |  |
| --- | --- |
| Attribute Name in FRANCHISE\_500 | Description |
| 2016\_rank | * Data Type: Qualitative (numeric) * Unique and Distinct value: 500 (no duplicates) * Data is clean. This is an ID column and also represented the ranks of the companies. |
| Company | * Data Type: Qualitative * Unique and Distinct value: 500 (no duplicates) * Data is clean. |
| Category\_detail | * Data Type: Qualitative * Unique value: 372 * Distinct value: 413 * *Data is dirty – lack of standardization of categories. However, this is not the priority at this stage.* |
| Franchise\_500\_rank | * Data Type: Qualitative (numeric) * Unique and Distinct value: 500 (no duplicates) * Data is clean. Same information as attribute ‘2016 Rank’. |
| Initial\_investment | * Data Type: Quantitative * Unique value: 486 * Distinct value: 493 * Data is clean. |
| High | * Data Type: Quantitative * Missing value: 1 * *Data is dirty – missing value.* |
| Change\_in\_units\_1\_yr | * Data Type: Qualitative * Missing values: 134 * Unique value: 339 * Distinct value: 353 * *Data is dirty – one value with different format at row 459 and missing values.* |
| Change\_in\_units\_3\_yr | * Data Type: Qualitative * Missing values: 117 * Unique value: 375 * Distinct value: 380 * *Data is dirty – one value with different format at row 459 and missing values.* |
| Founded | * Data Type: Quantitative * Data is clean |
| franchising\_since | * Data Type: Quantitative * Data is clean |
| corporate\_address | * Data Type: Qualitative * Data is clean |
| veteran incentives | * Data Type: Qualitative * Missing value: 152 * Unique value: 85 * Distinct value: 107 * *Data is dirty – missing values.* |
| highest rank | * Data Type: Quantitative * Missing value: 132 * Unique value: 134 * Distinct value: 225 * *Data is dirty – missing values.* |
| Lowest rank | * Data Type: Quantitative * Missing value: 132 * Unique value: 144 * Distinct value:235 * *Data is dirty – missing values.* |
| Where\_seeking\_franchise | * Data Type: Qualitative * Missing value: 128 (replaced with ‘unknown’) * Distinct value: 7 (should be 3) * *Data is dirty – mislabelled classes (eg. ‘Worldwide. ‘ vs ‘Worldwide’)* |
| operations | * Data Type: Qualitative * Missing values: 9 * Distinct value: 117 * Unique value: 70 * *Data is dirty – missing values and lack of standardization of categories.* |

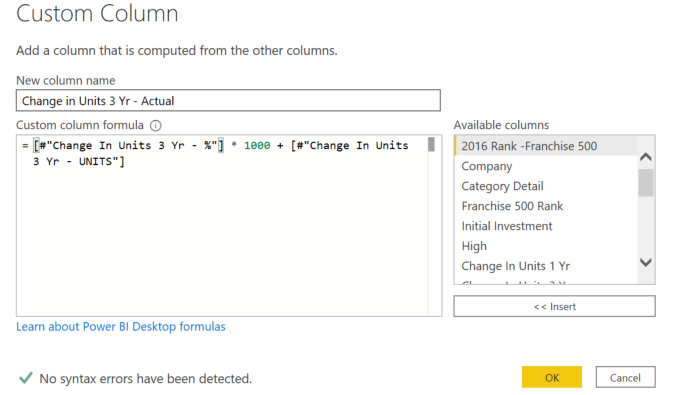
|  |  |
| --- | --- |
| **Attribute Name in FRANCHISE\_CEO** | **Description** |
| **company** | * Data Type: Qualitative * Unique and Distinct value: 500 (no duplicates) * Data is clean. This will be the ‘join key’ with FRANCHISE\_500. |
| **Ceo** | * Data Type: Qualitative * Data is clean. |
| **Parent company** | * Data Type: Qualitative * Missing value: 246 * ***Data is dirty – missing values*** |
| **net worth required** | * Data Type: Quantitative * Missing value: 94 * Distinct value: 116 * Unique value: 80 * ***Data is dirty – missing values and inconsistent format of values, i.e some are single values and some values are in range format.*** |
| **Liquid cash requirement** | * Data Type: Quantitative * Missing value: 52 * Distinct value: 152 * Unique value: 107 * ***Data is dirty – missing values and inconsistent format of values, i.e some are single values and some values are in range format.***      * ***range.*** |
| **Initial franchise fee** | * Data Type: Quantitative * Missing value: 6 * ***Data is dirty – missing values*** |

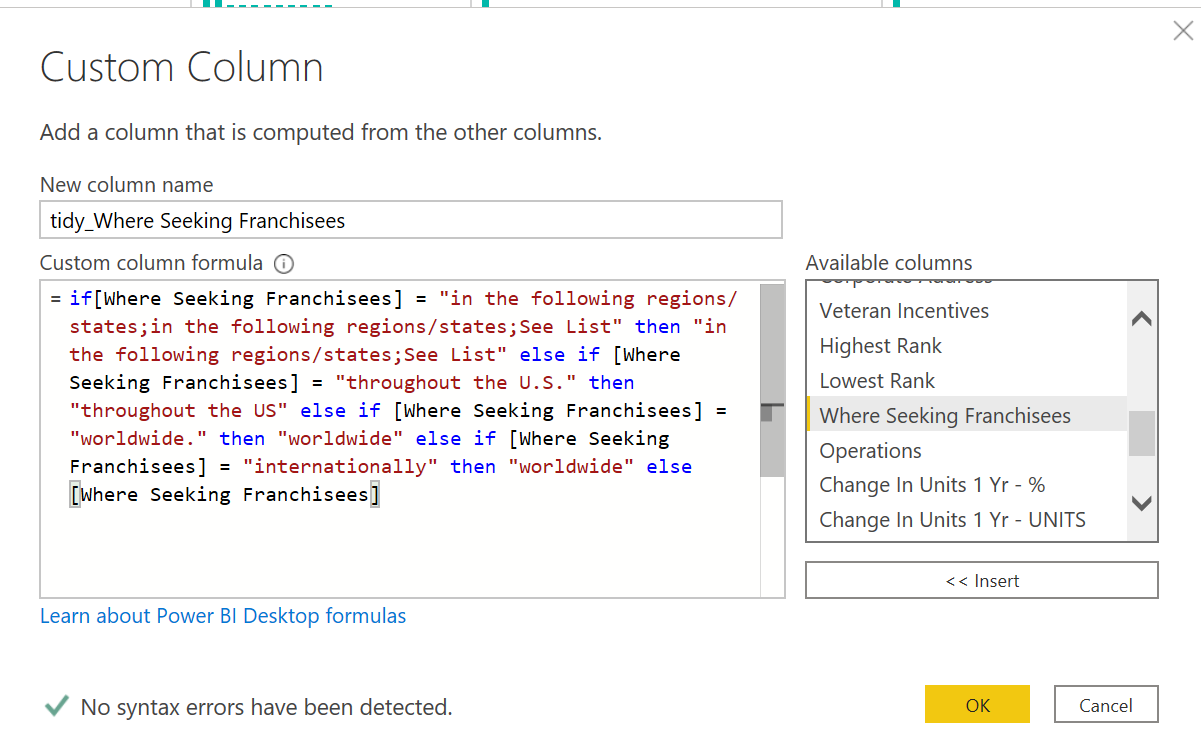
**Annex B – Data Profiling on all attributes of FRANCHISE\_500 and FRANCHISE\_CEO**

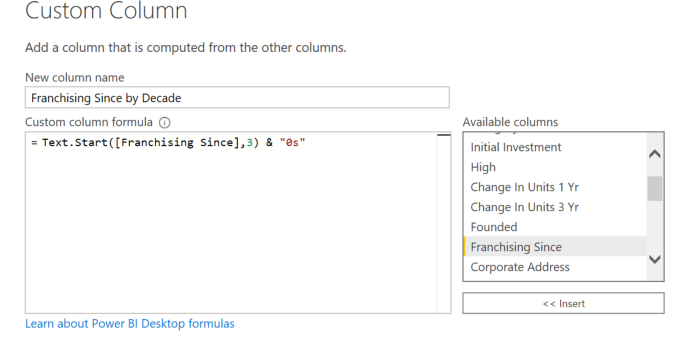
|  |  |
| --- | --- |
| Attribute Name in FRANCHISE\_500 | Description |
| 2016\_rank | Ranking of Franchise in 2016 |
| company | Name of company |
| category\_detail | Business category |
| franchise\_500\_rank | Top 500 Franchise ranking |
| initial\_investment | Initial investment required for franchising |
| high | Maximum amount of investment required for franchising |
| change\_in\_units\_1\_yr | Change in franchise units within 1 year |
| change\_in\_units\_3\_yr | Change in franchise units within 3 year |
| founded | Year the company is founded |
| franchising\_since | Year Franchising started |
| corporate\_address | Corporate Address |
| veteran incentives | Incentives given to franchisee operators |
| highest rank | Highest ranking on the franchise listing |
| lowest rank | Lowest ranking on the franchise listing |
| where\_seeking\_franchise | Locations where Franchising is available |
| operations | The type of training provided for franchising |

**Annex C – Screenshot of formulae used for calculated columns**

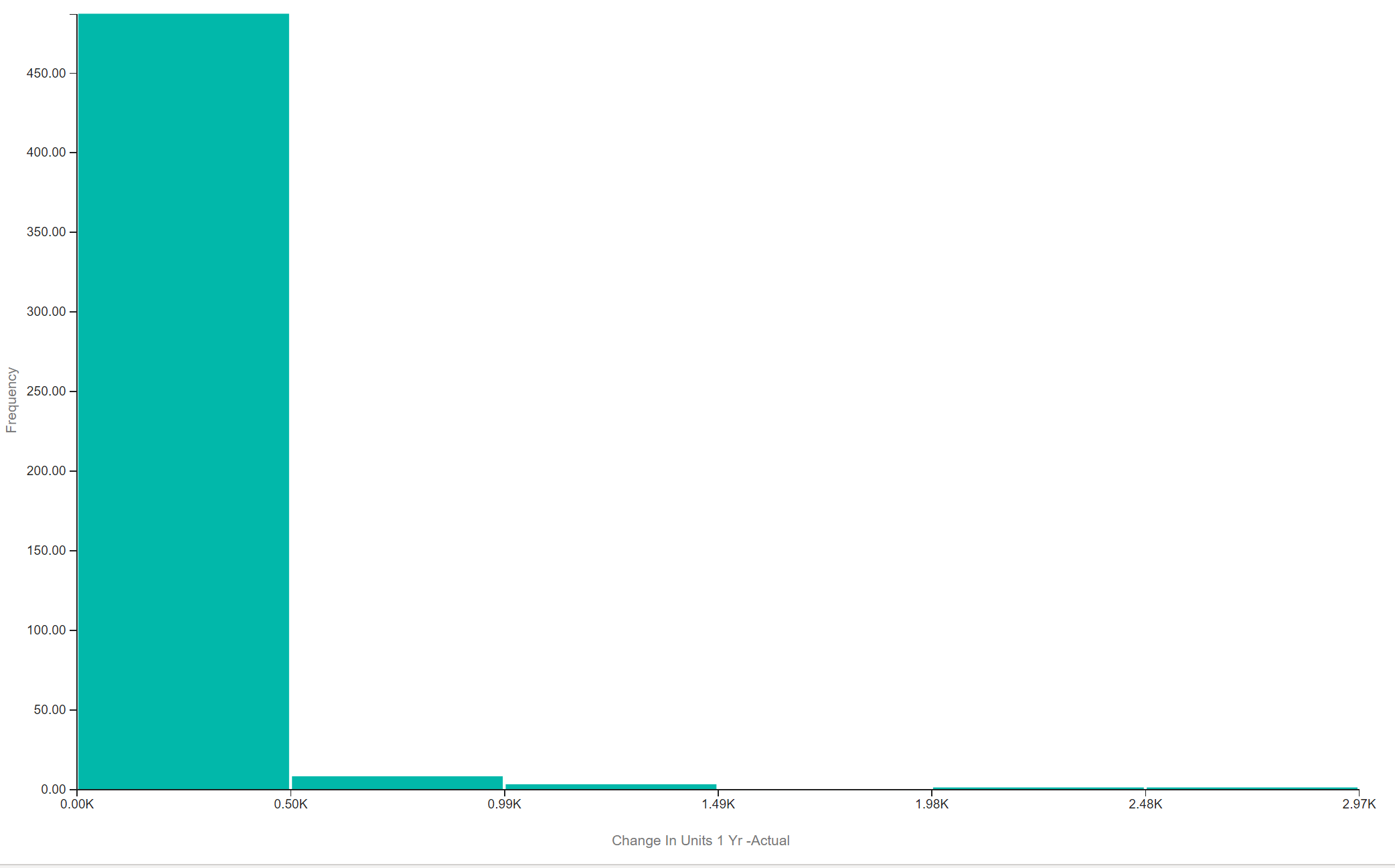






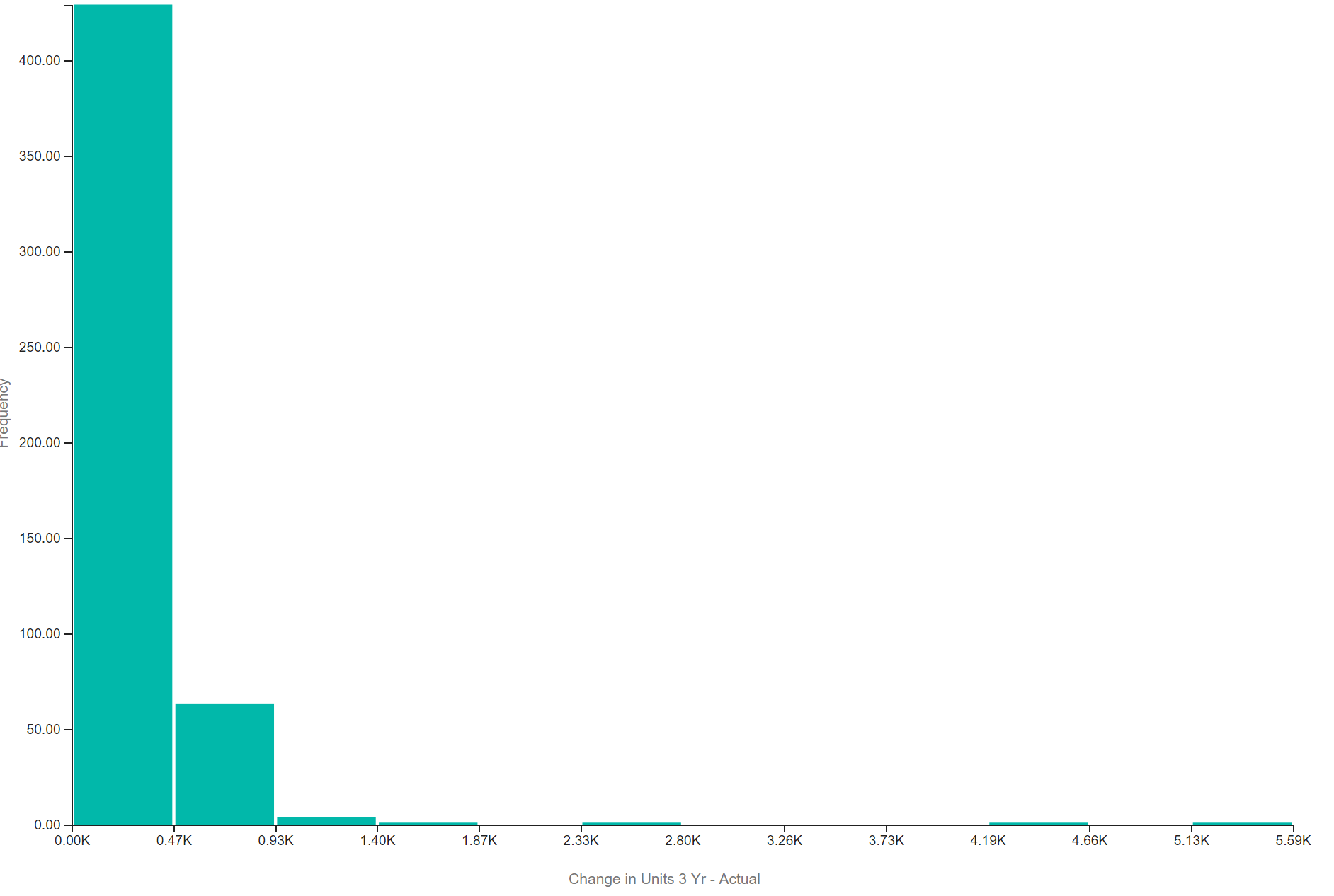


**Annex D: Histograms for Change In Units 1 Yr, Change In Units 3 Yr, Liquid Cash Requirement and Net Worth Requirement**



Outliers

Outliers

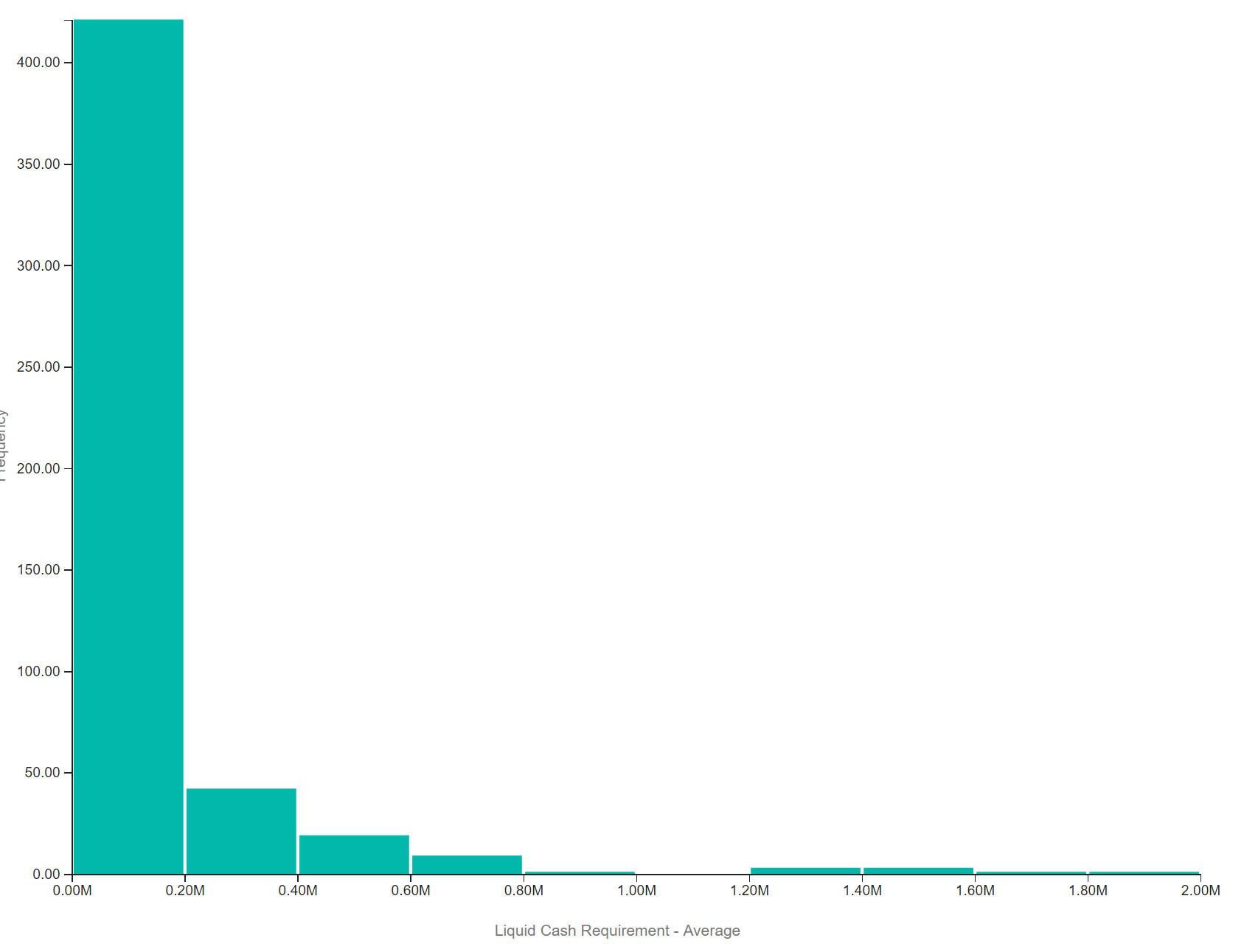


outlier

Outliers

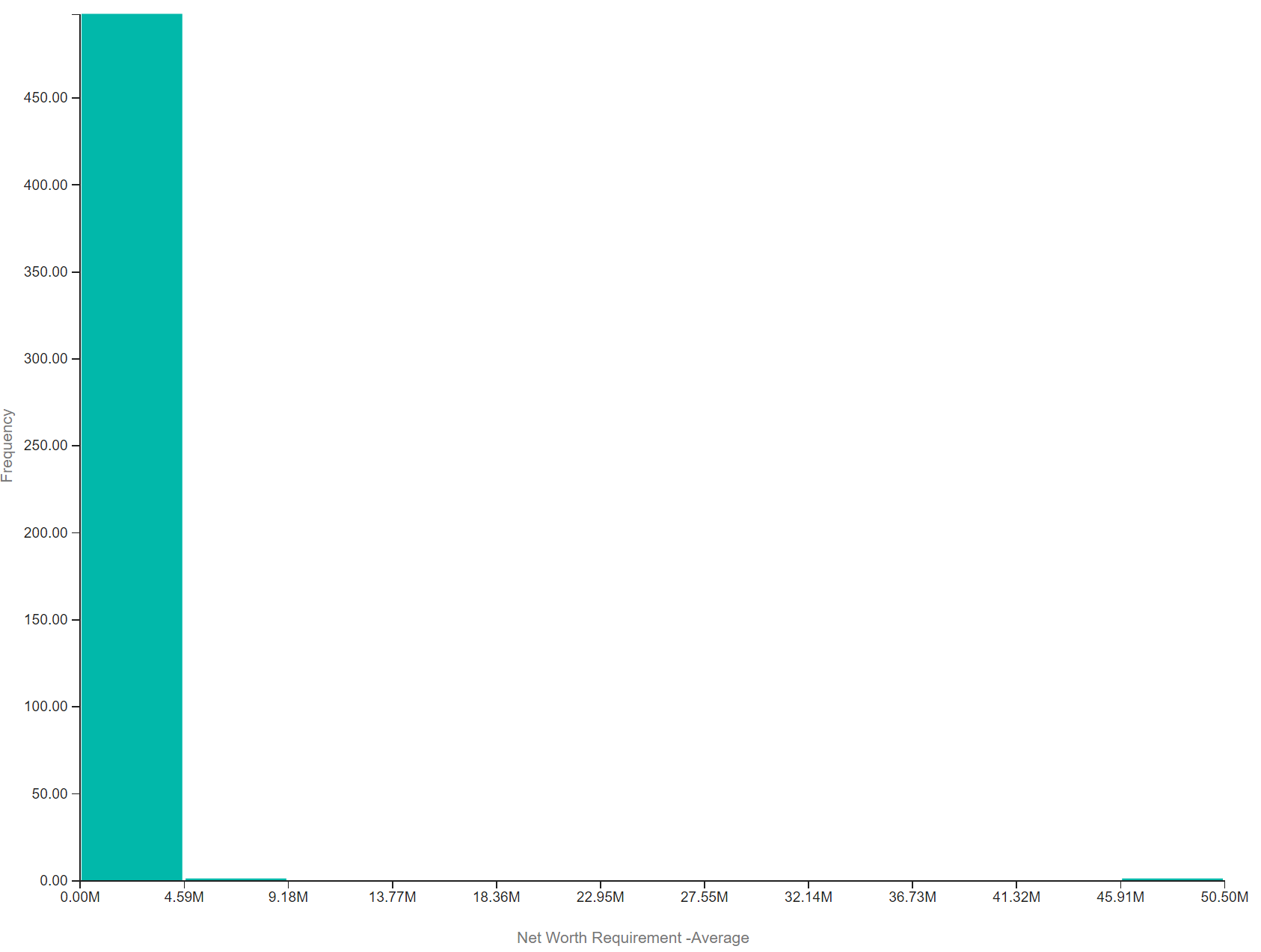
Outliers

Outlier



Outliers

Outliers



Outlier

Outliers