Team Nintendo DS - Final Report

Richie Youm, Mena Ayad, Dekan Zeng, Rahul Malhotra

Introduction

Venture Inc., a company which is new to the cannabis industry, has received regulatory approval to open retail locations in the Toronto Census Metropolitan Area to sell cannabis. In addition to Venture Inc., there are two other companies which obtains similar rights for the area to the west and east of the Toronto CMA. With its special position in the marketplace, Venture Inc. plans to strategically begin their venture and also has the intention to acquire these companies to create a monopoly position in the market. In order to be competitive in the acquisition negotiating process and increase probability of success, the following market analysis have been conducted based on the Toronto CMA data and the databases provided by Environic Analytics.

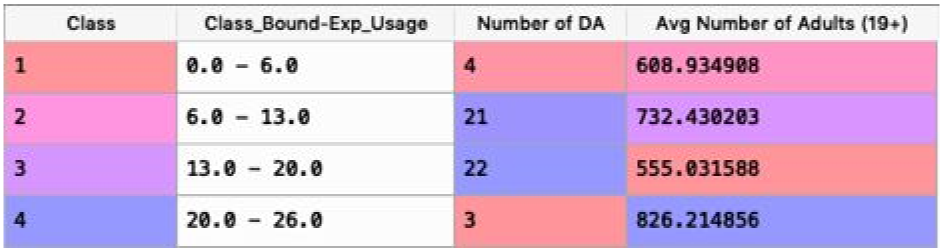
Data audit

A data audit was conducted on all of our model predictors and the target variable. There was no NA value in the dataset, there were irrelevant rows of observations called "UN", which were removed from the dataset.

The descriptive statistics of the target variable 'DEPVAR7', cannabis consumption per person aged 19+ (in $/mth) was analyzed, in which had a mean of about $11.4 and mean of $10 (Appx 1). In general, we would expect a successful store to have at least $11.4 of expected usage by customer. While standard deviation was considerably large at $5.8, we did not discover any other concerning problems with the target variable.

Sparsity was also checked during the audit. The variable PRIZM5, which contains the Dissemination Area (DA) labels, was dummified as binary columns, which resulted in a sparse data. For example, PRIZM5 Segment Code 20 (PRIZM5DA_20) only consisted of 4.35% of the total population (Appx 2). However, those variables were expected to be sparse and it was a very significant variable in the regression model, so we decided to tolerate its sparsity. Overall, most of top significant independent variables were normally or very close to normally distributed, without any other concerning problems.

Deliverable 1: Frequency distribution of classes of expected cannabis usage



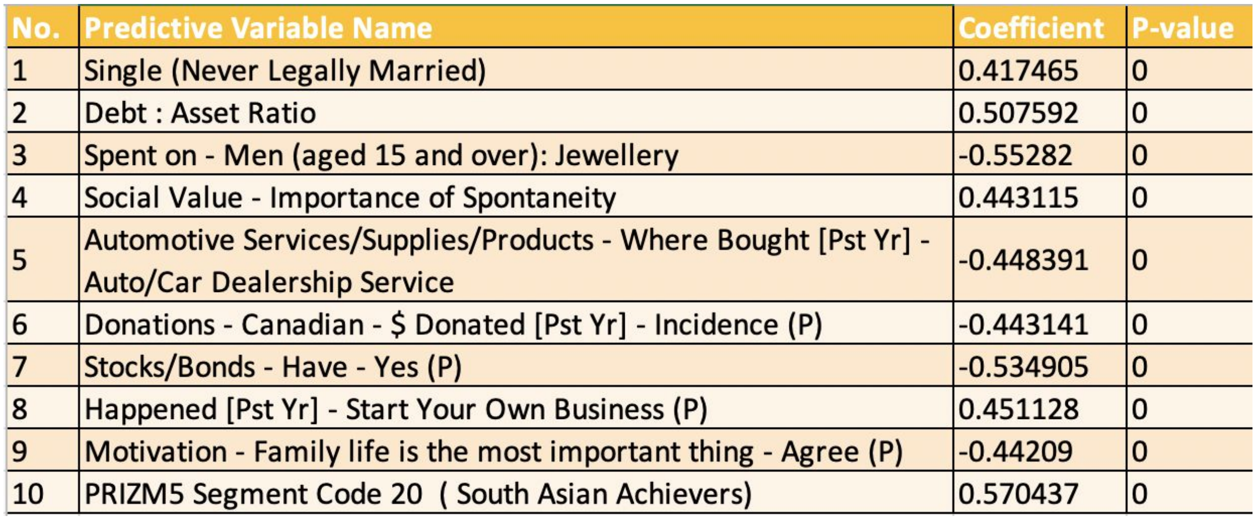
Class	Class_Bound-Exp_Usage	Number of DA	Avg Number of Adults (19+)
1	0.0 - 6.0	4	608.934908
2	6.0 - 13.0	21	732.430203
3	13.0 - 20.0	22	555.031588
4	20.0 - 26.0	3	826.214856

The frequency distribution table contains 4 classes, in which labels their respective range of average cannabis usage per person. The majority of DAs falls into class 2 and class 3 equally. However, the average number of Adults (19+) in class 2 is significantly higher than in class 3.

Process of splitting "Development file" and "Validation file"

The analytical file above has been split into two equal parts for further analysis. We chose to divide the main dataset using sklearn package called "train_test_split". We used a test size of 50% as instructed, and set a random state in order to reproduce consistent result, thus achieving a semi-random splits.

Deliverable 2: Pearson product moment correlations - Top 10 correlated variables



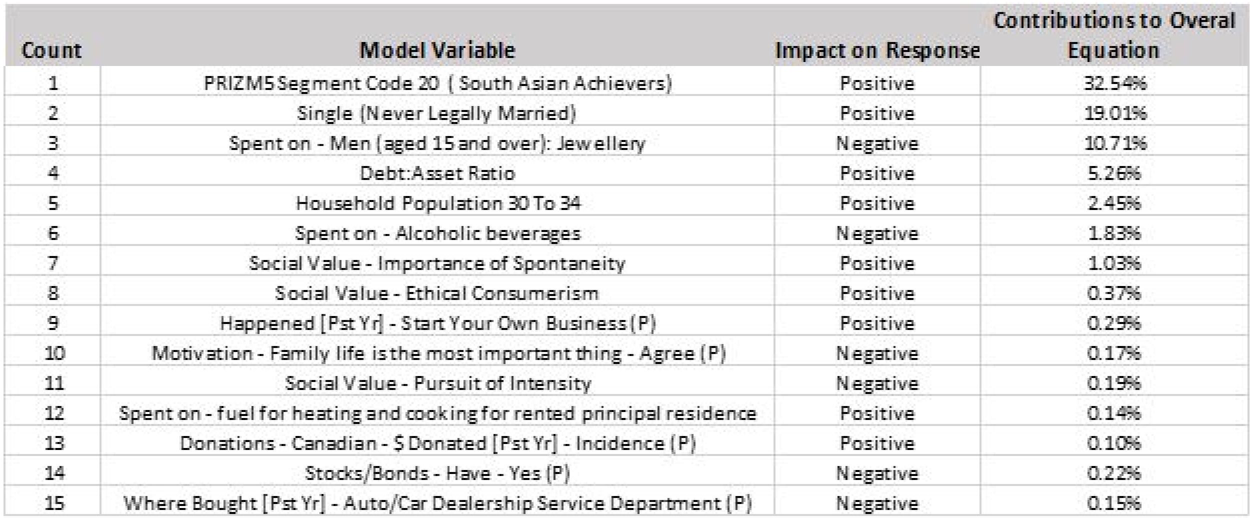
No.	Predictive Variable Name	Coefficient	P-value
1	Single (Never Legally Married)	0.417465	0
2	Debt : Asset Ratio	0.507592	0
3	Spent on - Men (aged 15 and over): Jewellery	-0.55282	0
4	Social Value - Importance of Spontaneity	0.443115	0
5	Automotive Services/Supplies/Products - Where Bought [Pst Yr] - Auto/Car Dealership Service	-0.448391	0
6	Donations - Canadian - $ Donated [Pst Yr] - Incidence (P)	-0.443141	0
7	Stocks/Bonds - Have - Yes (P)	-0.534905	0
8	Happened [Pst Yr] - Start Your Own Business (P)	0.451128	0
9	Motivation - Family life is the most important thing - Agree (P)	-0.44209	0
10	PRIZM5 Segment Code 20 (South Asian Achievers)	0.570437	0

The correlation between the target and all reasonable independent variables has been examined, and the output was ranked from the most significant to the least significant. Subsequently, the top 10 correlated variables in absolute values with a statistical significance of at least 99% were chosen and demonstrated above.

Process of removing multicollinearity

If there are multicollinear pairs, we simply need to remove one of them. Since there are many features in the dataset, we first appended a list of multicollinear pairs and filtered our correlation chart by the list. Then, loop was constructed so that a column is removed and checked if the column contained any correlation greater than threshold of .65 (absolute value), repeated until we there wasn't any multicollinear pairs. Eventually, the multicollinear variables, which was the difference between the originally selected multicollinear variables and what was left out of the loop, was filtered out of the main data.

Deliverable 3: final model variables report



Count	Model Variable	Impact on Response	Contributions to Overal Equation
1	PRIZM5 Segment Code 20 (South Asian Achievers)	Positive	32.54%
2	Single (Never Legally Married)	Positive	19.01%
3	Spent on - Men (aged 15 and over): Jewellery	Negative	10.71%
4	Debt:Asset Ratio	Positive	5.26%
5	Household Population 30 To 34	Positive	2.45%
6	Spent on - Alcoholic beverages	Negative	1.83%
7	Social Value - Importance of Spontaneity	Positive	1.03%
8	Social Value - Ethical Consumerism	Positive	0.37%
9	Happened [Pst Yr] - Start Your Own Business (P)	Positive	0.29%
10	Motivation - Family life is the most important thing - Agree (P)	Negative	0.17%
11	Social Value - Pursuit of Intensity	Negative	0.19%
12	Spent on - fuel for heating and cooking for rented principal residence	Positive	0.14%
13	Donations - Canadian - $ Donated [Pst Yr] - Incidence (P)	Positive	0.10%
14	Stocks/Bonds - Have - Yes (P)	Negative	0.22%
15	Where Bought [Pst Yr] - Auto/Car Dealership Service Department (P)	Negative	0.15%

Additional observations of the top 15 variables

Those that are associated with "PRIZM5DA_20" also spend on home entertainment equipment and services, computer hardware, and accessories for men (aged 15 and over), as they all have strong positive correlation.

❑ This signals for an opportunity for online sales which could boost sales as greater spending in computer hardware suggests more presence in the internet. Those customers are more likely to be familiar with online shopping, in which they may also have the tendency to purchase products online rather than going to the brick-and-mortar stores.

Those that are single belongs mostly to the household population of 25 ~ 29. In addition, they have a strong positive correlation with having social value of believing in the pursuit of intensity, valuing aesthetics and intuition & impulse.

❑ This shows the possibility of cooperation with extreme sports brand will increase the sales of the cannabis.

Those that value ethical consumerism have strong positive correlation with having the social value of personal creativity.

❑ Cannabis is often associated with freedom, due to its cultural background of being an illegal, restricted substance. We see this as a part of the reason for such correlation between those variables

❑ This provides a better picture of the profiles of some of our potential consumers

People who spend on alcoholic beverages are much less likely to spend their money on entertainment, tuition fees, and cigarettes.

❑ This suggests that alcohol is considered a substitute for cannabis.

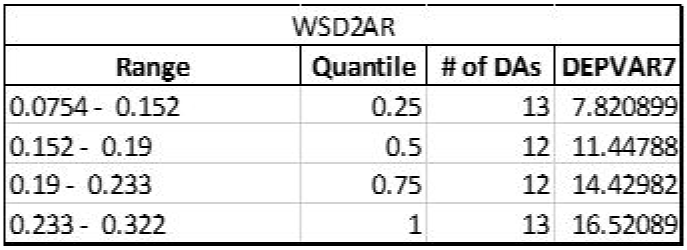
❑ Maybe, a promotion contrasting the consequences of alcohol consumption which are already known to be severe, with the cannabis consumption which are less associated with such consequences, could serve as a benefit in people's recognition of cannabis

Deliverable 4: Exploratory Data Analysis (EDA) report

Examining the trend of some of the independent variables versus the target variables:

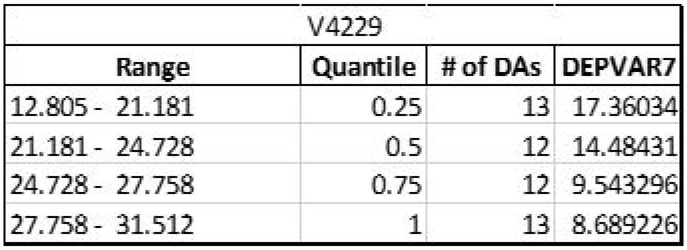
- Debt: Asset Ratio (WSD2AR)

There is a strong positive trend relation WSD2AR and the target variable, the higher the ratio of debt to assets, the more the consumers spend on the cannabis monthly. The relation is so strong that customers in the last quartile, by debt to assets ratio, spends double what the customers on the 1st quartile do.



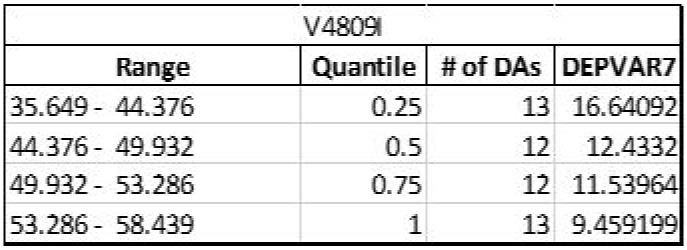
WSD2AR			
Range	**Quantile**	**# of DAs**	**DEPVAR7**
0.0754 - 0.152	0.25	13	7.820899
0.152 - 0.19	0.5	12	11.44788
0.19 - 0.233	0.75	12	14.42982
0.233 - 0.322	1	13	16.52089

- Automotive Services/Supplies/Products - Where Bought [Pst Yr] - Auto/Car Dealership Service Department (P) (V4229): Interestingly unlike (WSD2AR), V4229 has a strong negative correlation with the monthly spending on cannabis.

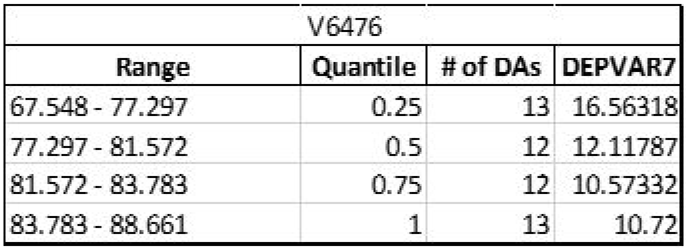


V4229			
Range	Quantile	# of DAs	DEPVAR7
12.805 - 21.181	0.25	13	17.36034
21.181 - 24.728	0.5	12	14.48431
24.728 - 27.758	0.75	12	9.543296
27.758 - 31.512	1	13	8.689226

- The same previous negative relation holds true too for Donations - Canadian - $ Donated [Pst Yr] - Incidence (P) (V4809I) variable and Motivation - Family life is the most important thing - Agree (P) (V6476). It can be concluded that the more socially involved consumers are less likely to spend on cannabis.

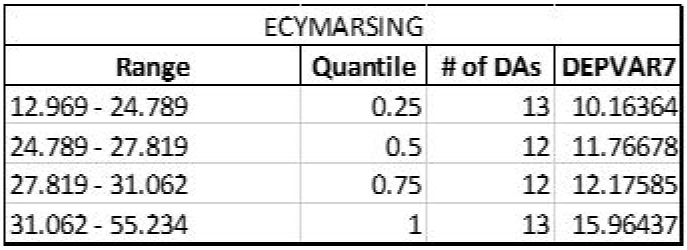


V4809I			
Range	Quantile	# of DAs	DEPVAR7
35.649 - 44.376	0.25	13	16.64092
44.376 - 49.932	0.5	12	12.4332
49.932 - 53.286	0.75	12	11.53964
53.286 - 58.439	1	13	9.459199

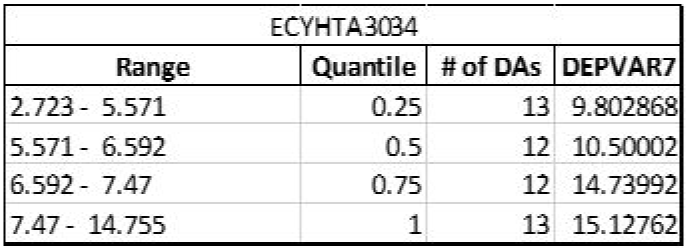


V6476			
Range	Quantile	# of DAs	DEPVAR7
67.548 - 77.297	0.25	13	16.56318
77.297 - 81.572	0.5	12	12.11787
81.572 - 83.783	0.75	12	10.57332
83.783 - 88.661	1	13	10.72

- Single (Never Legally Married) (ECYMARSING) and Household Population 30 To 34 (ECYHTA3034) variable have strongly positive trend relation with our target variable. Being single and aged in late 20s to late 30s play an important role in spending on cannabis.

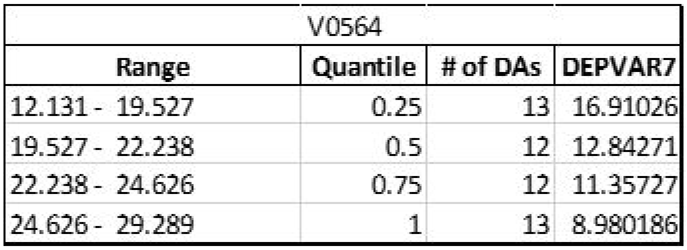


ECYMARSING			
Range	Quantile	# of DAs	DEPVAR7
12.969 - 24.789	0.25	13	10.16364
24.789 - 27.819	0.5	12	11.76678
27.819 - 31.062	0.75	12	12.17585
31.062 - 55.234	1	13	15.96437

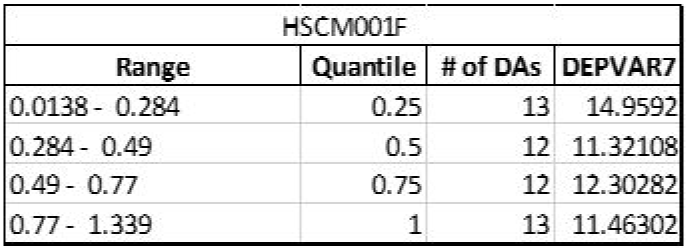


ECYHTA3034			
Range	Quantile	# of DAs	DEPVAR7
2.723 - 5.571	0.25	13	9.802868
5.571 - 6.592	0.5	12	10.50002
6.592 - 7.47	0.75	12	14.73992
7.47 - 14.755	1	13	15.12762

- Stocks/Bonds - Have - Yes (P) (V0564) and Spent on - Men (aged 15 and over): Jewellry (HSCM001F) have a strong negative trended with spending on cannabis. The consumers who invested $24.626 to $29.289 monthly spend on cannabis on average half what the consumers who invested $12.131 to $19.527 only.



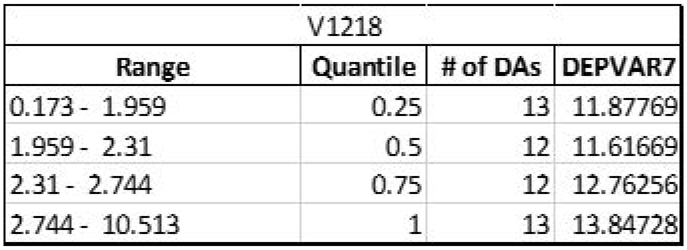
V0564			
Range	Quantile	# of DAs	DEPVAR7
12.131 - 19.527	0.25	13	16.91026
19.527 - 22.238	0.5	12	12.84271
22.238 - 24.626	0.75	12	11.35727
24.626 - 29.289	1	13	8.980186



HSCM001F			
Range	Quantile	# of DAs	DEPVAR7
0.0138 - 0.284	0.25	13	14.9592
0.284 - 0.49	0.5	12	11.32108
0.49 - 0.77	0.75	12	12.30282
0.77 - 1.339	1	13	11.46302

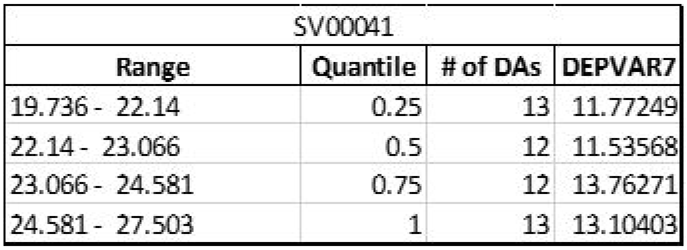
The EDA of rest of independent variables:

- Happened [Pst Yr] - Start Your Own Business (P) (V1218) variable



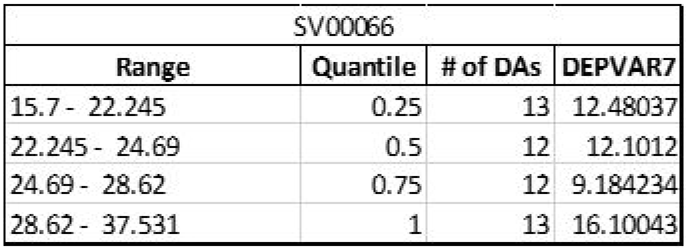
V1218			
Range	Quantile	# of DAs	DEPVAR7
0.173 - 1.959	0.25	13	11.87769
1.959 - 2.31	0.5	12	11.61669
2.31 - 2.744	0.75	12	12.76256
2.744 - 10.513	1	13	13.84728

- Social Value - Importance of Spontaneity (SV00041) variable



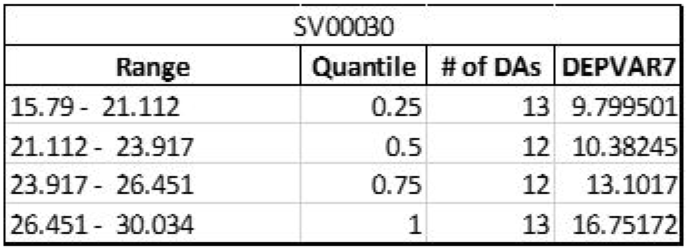
SV00041			
Range	Quantile	# of DAs	DEPVAR7
19.736 - 22.14	0.25	13	11.77249
22.14 - 23.066	0.5	12	11.53568
23.066 - 24.581	0.75	12	13.76271
24.581 - 27.503	1	13	13.10403

- Social Value - Pursuit of Intensity (SV00066) variable



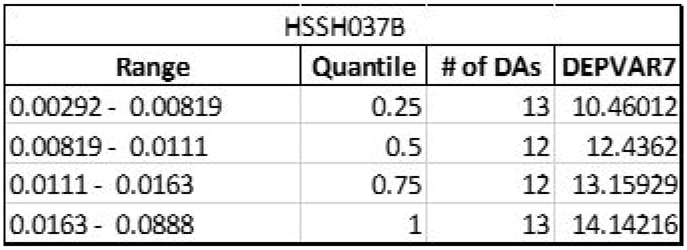
SV00066			
Range	Quantile	# of DAs	DEPVAR7
15.7 - 22.245	0.25	13	12.48037
22.245 - 24.69	0.5	12	12.1012
24.69 - 28.62	0.75	12	9.184234
28.62 - 37.531	1	13	16.10043

- Social Value - Ethical Consumerism (SV00030) variable



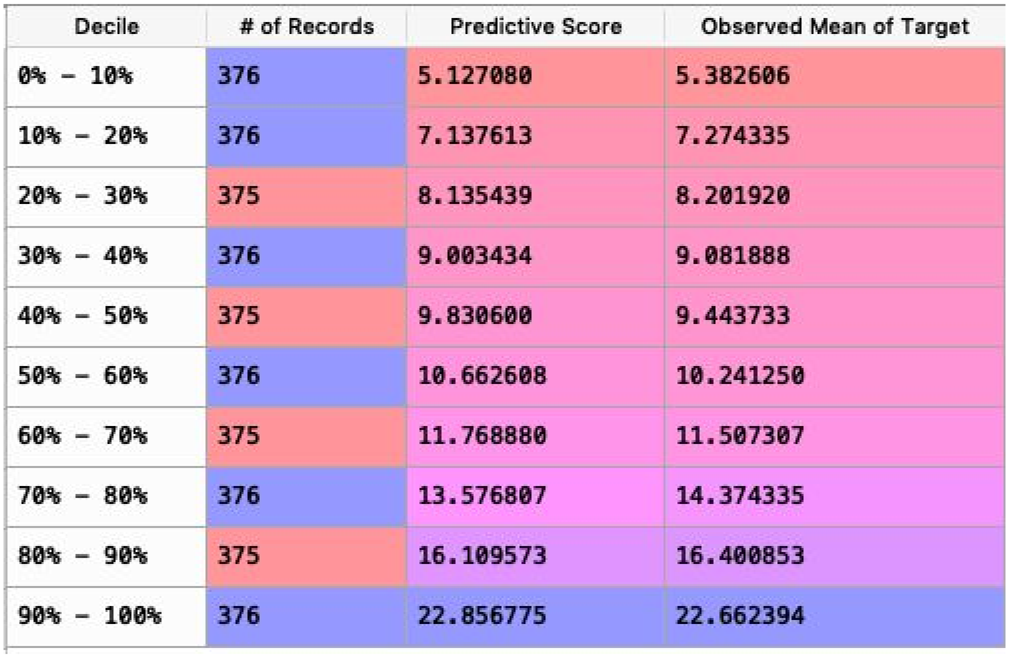
SV00030			
Range	Quantile	# of DAs	DEPVAR7
15.79 - 21.112	0.25	13	9.799501
21.112 - 23.917	0.5	12	10.38245
23.917 - 26.451	0.75	12	13.1017
26.451 - 30.034	1	13	16.75172

- Spent on - Wood and other fuel for heating and cooking for rented principal residence (HSSH037B) variable



HSSH037B			
Range	Quantile	# of DAs	DEPVAR7
0.00292 - 0.00819	0.25	13	10.46012
0.00819 - 0.0111	0.5	12	12.4362
0.0111 - 0.0163	0.75	12	13.15929
0.0163 - 0.0888	1	13	14.14216

Deliverable 5: Decile report



Decile	# of Records	Predictive Score	Observed Mean of Target
0% - 10%	376	5.127080	5.382606
10% - 20%	376	7.137613	7.274335
20% - 30%	375	8.135439	8.201920
30% - 40%	376	9.003434	9.081888
40% - 50%	375	9.830600	9.443733
50% - 60%	376	10.662608	10.241250
60% - 70%	375	11.768880	11.507307
70% - 80%	376	13.576807	14.374335
80% - 90%	375	16.109573	16.400853
90% - 100%	376	22.856775	22.662394

Deliverable 6: Description of overall approach and methodology

Preprocessing:

Most of the steps mentioned in the guideline have been followed with a few deviations for better solutions. Initially, data auditing and exploration have been conducted. We were interested in observing potential impact of having a young child in the household, say, less than 10 years old, and created a dummy variable to determine whether this would impact the expected cannabis usage. In order to integrate the Prizm5 into our model, we also created dummy variables for the DAs assigned by Prizm5. There are 50 unique DAs in our datasets, and in order to avoid “dummy variable trap”, we excluded one dummy variable from the list and used it as a base value.

Model building:

The correlations between the target variable and all independent variables have been examined and ranked from the most significant to least significant. Those variables with a statistical significance greater than 99 % have been chosen to append a list of the top 20 variables which were to be used in the model. As there are many features in the dataset, independent variable set could have been strongly related. Therefore, there was a necessity to take the multicollinearity into consideration during the analysis process. A verification has been carried out to identify those pairs that are multicollinear (greater than 65% of correlation, in absolute term) and one of the identified pair was removed from the variable set, which was repeated until there was no pairs left.

In addition, a stepwise regression on the top 20 correlated variables was carried out with the consideration of multicollinearity. We assumed to keep the same requirement for p-value at 0.01, which resulted in the removal of 5 variables leaving us with 15 variables in the model.

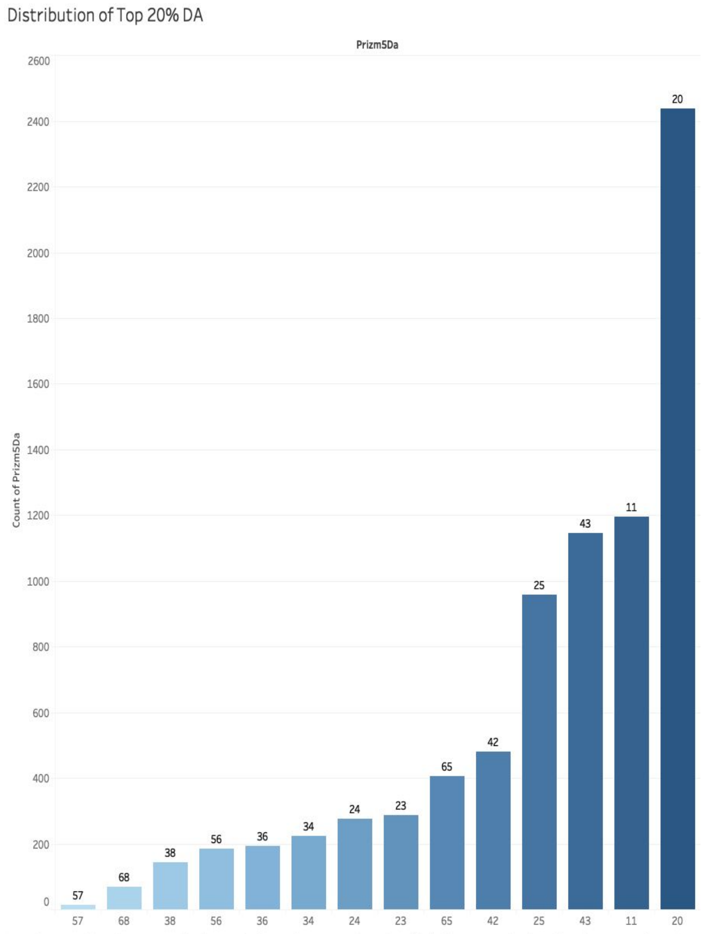
After having the model with most significant 15 variables, the dataset was divided 50-50 into training and validation file. The reason why we chose to move this step afterwards is because we did not want to lose any correlation of data from the splitted development file; therefore, we used the entire dataset to determine correlation and multicollinearity to get a better picture of the general relationship between variables.

To depict the trend of the independent variable versus the dependent variable, an Exploratory Data Analysis report for all final model predictive variables was produced. The EDA report focuses on the relationship between our 15 selected variables and the target variable.

Once the final predictive model was set up, a validation process was implemented. By simply applying the model equation to the validation file, we sorted the validation sample by model score into 10 deciles (with the top decile having the highest model scores and bottom decile having the lowest model scores) and prepared a table as in the format as requested.

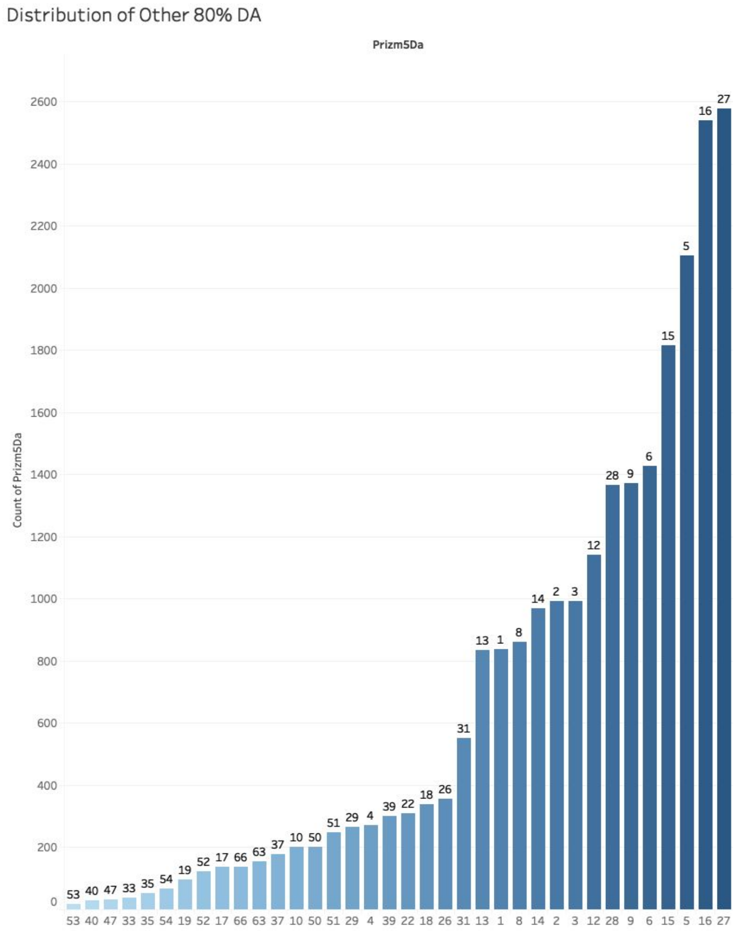
The result of our final model was quite impressive; the predictions had MSE of 8.78, in which did not fall much short from more sophisticated machine learning models such as such as Random Forest. From the sheer accuracy of our prediction, we are reasonably confident that this model could be adjusted and applied, and extrapolated to regions outside of CMA. This would serve as a negotiating tool in acquiring the other two companies that Venture Inc. intends to control in the future.

Deliverable 7: Distribution of PRIZM5 segments (DA) and Lifestage Groups (LS); Top 20% vs Remaining 80%



Distribution of Top 20% DA

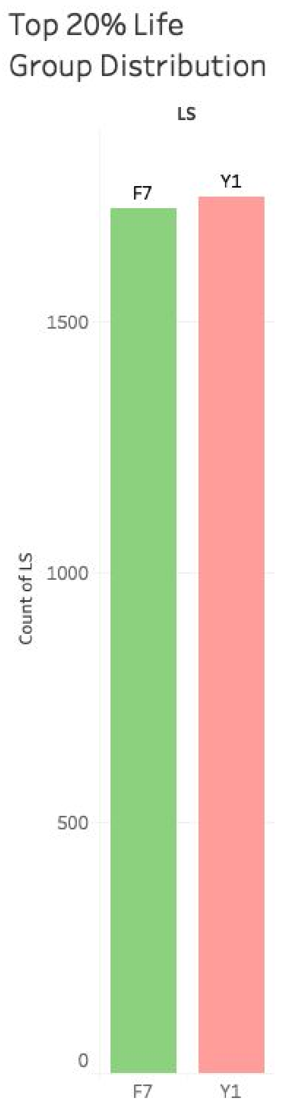
Count of Prizm5Da for each Prizm5Da. Color shows Rank of Count of Prizm5Da. The marks are labeled by Prizm5Da. The data is filtered on average of Top DA, which includes values greater than or equal to 1.



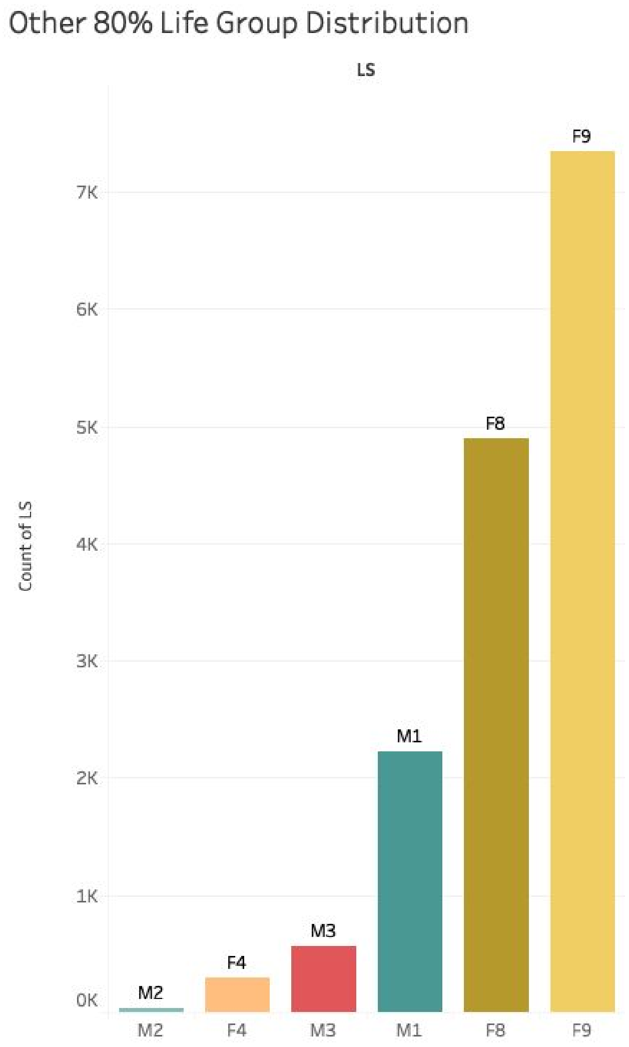
Distribution of Other 80% DA

Count of Prizm5Da for each Prizm5Da. Color shows Rank of Count of Prizm5Da. The marks are labeled by Prizm5Da. The data is filtered on average of Bottom DA, which includes values greater than or equal to 1.

The above graphs represent the distribution of DAs in the top 20% and remaining 80% of DAs. As seen, DA 20 has the most significant representation within the top group by a significant margin, whereas DA 16 and 27 showed greatest representation within the remainder group. These findings indicate that if we are able to effectively capture the demand of DA 20 which composes the majority of the top group, we could maximize our revenue. A corresponding evidence can be found in expected cannabis usages; DA 20, having the most amount of representation, also has the highest expected usage per person (Appx 3). On the other hand, DA 16 and 27 have a considerably lower expected usage, at about $8 and $10.5, respectively. Despite expected usage being lower, it is too early to conclude that it would be a bad idea to open stores near these DA in the future from this finding. Rather, a fair conclusion from this observation is that stores near those DAs are likely to have a lower expected usage per person. Therefore, if possible, we should consider opening stores within clusters of DA 20 or at the clusters of following DAs.


Top 20% Life Group Distribution

Count of LS for each LS. Color shows details about LS. The marks are labeled by LS. The data is filtered on average of Top DA, which includes values greater than or equal to 1.


Other 80% Life Group Distribution

Count of LS for each LS. Color shows details about LS. The marks are labeled by LS. The data is filtered on average of Bottom DA, which includes values greater than or equal to 1.

Life group distribution was much simpler. In the top social group, F7 and Y1 had an even distribution whereas F9 composed of the majority in the remainder group. Interestingly, majority of the CMA population lies in the F8 and F9 life stage. While exact meaning of each code is unclear due to lack of information given, nonetheless, our focus is on the top 20% of DAs. This indicates that people in those stage are the heaviest users of cannabis, in which we could further explore opportunities for future growth and strategic planning.

Deliverable 8: Proposed venture strategy

1. Store opening:

So far, we have successfully identified the characteristics of the most profitable cannabis customers, locations with strong demand, as well as some unfavorable locations with weaker demand. In the first map below represents the distribution of DAs in the top 20% group, in which we can see clusters of DAs with one of the highest expected average usages, namely DA 20, DA 34, and DA 11. (Map 1)

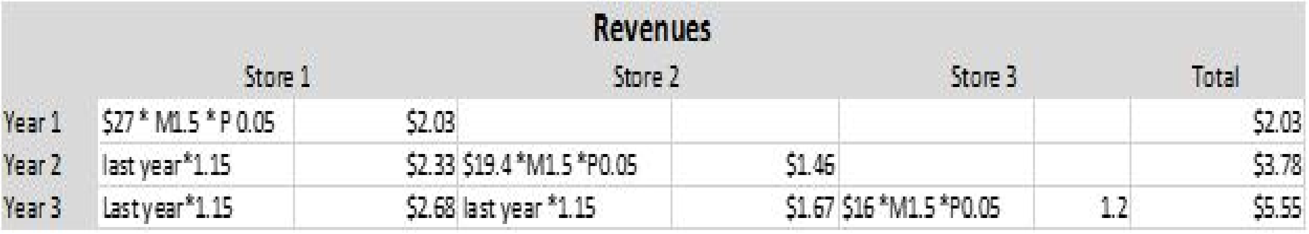
The second map is a comparative visualization of the locations of the DAs and populations 19+ that distinguishes distribution between top 20% DAs (Blue) and remaining 80% (Red). (Map 2) Ideally, we need to locate the stores in areas where there are DAs with strong demand and large population (large blue), and furthermore avoid locations that are predominantly populated with lower usage consumers (large red) or where the populations are too scattered.

Our primary goal is to capture the most amount of demand where there is highest profitability, and based on this criteria we have chosen three locations that can be most accessible to high-usage customers in the most populated area. (Map 3) At each locations, we are able to capture approximately 1.5 million customers with average usage of about $27, $19.4, $16 per person, respectively. Note, that our downtown Toronto location covers both DA 34 and DA 11, and for our financials, average between the two will be used.

The last map denotes potential store locations in the future. (Map 4) As seen, there are many sizeable opportunities mainly in the clusters of DA 20 in many areas, as well as a few areas in the downtown Toronto. The points were used as a proxy of the next most effective areas that could maximize our profits in the long run.

2. Financial projections based on population and average spending in targeted areas:

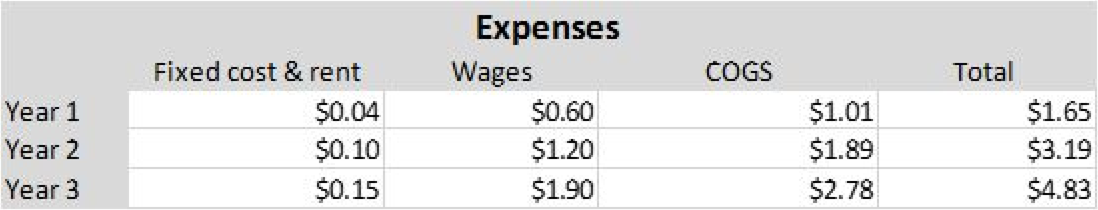
Revenue projections for the first 3 years:
- With assumptions of initial market penetration (P) of 5% of total population (M) in the target area and annual compounded growth rate of 15% .



Revenues							
	Store 1		Store 2		Store 3		Total
Year 1	$27 * M1.5 * P 0.05	$2.03					$2.03
Year 2	last year*1.15	$2.33	$19.4 *M1.5 *P0.05	$1.46			$3.78
Year 3	Lastyear*1.15	$2.68	last year *1.15	$1.67	$16 *M1.5 *P0.05	1.2	$5.55

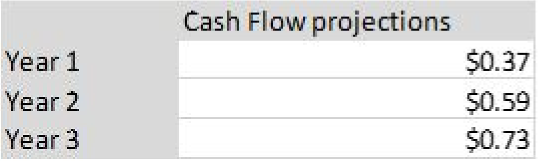
Expenses projections:

- With the assumption of 50% profit margin on the sold goods. Wages are calculated for 20 full time worker in the first year and grow as we add more stores.



Expenses				
	Fixed cost & rent	Wages	COGS	Total
Year 1	$0.04	$0.60	$1.01	$1.65
Year 2	$0.10	$1.20	$1.89	$3.19
Year 3	$0.15	$1.90	$2.78	$4.83

Expected Cash flow:



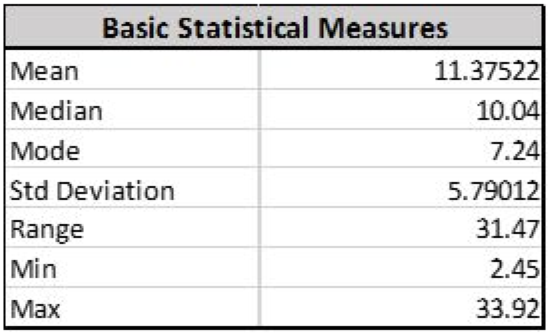
	Cash Flow projections
Year 1	$0.37
Year 2	$0.59
Year 3	$0.73

Notes

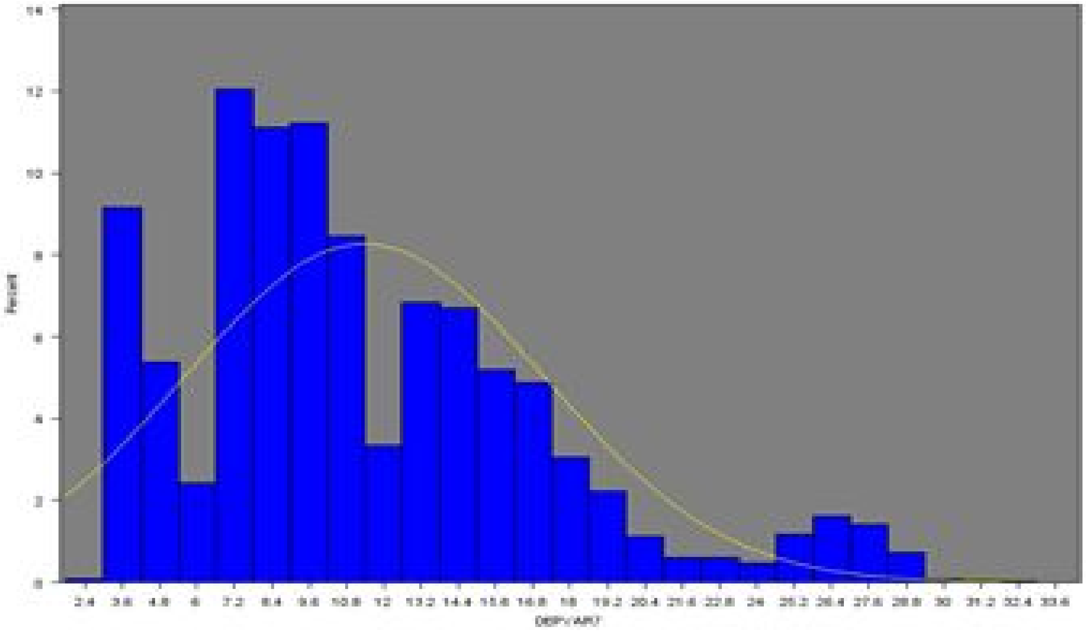
- With assumption of no market competitors, since exclusive retail license was given to Venture Inc.
- All the dollar figures are in terms of million dollar.

Appendix:-

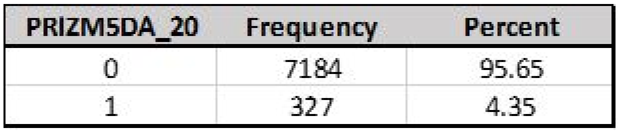
Appendix 1: Descriptive statistics for target variable & PRIZM5DA_20



Basic Statistical Measures	
Mean	11.37522
Median	10.04
Mode	7.24
Std Deviation	5.79012
Range	31.47
Min	2.45
Max	33.92

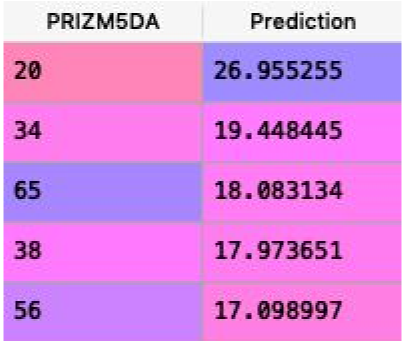


Appendix 2: Frequency chart of PRIZM5DA_20



PRIZM5DA_20	Frequency	Percent
0	7184	95.65
1	327	4.35

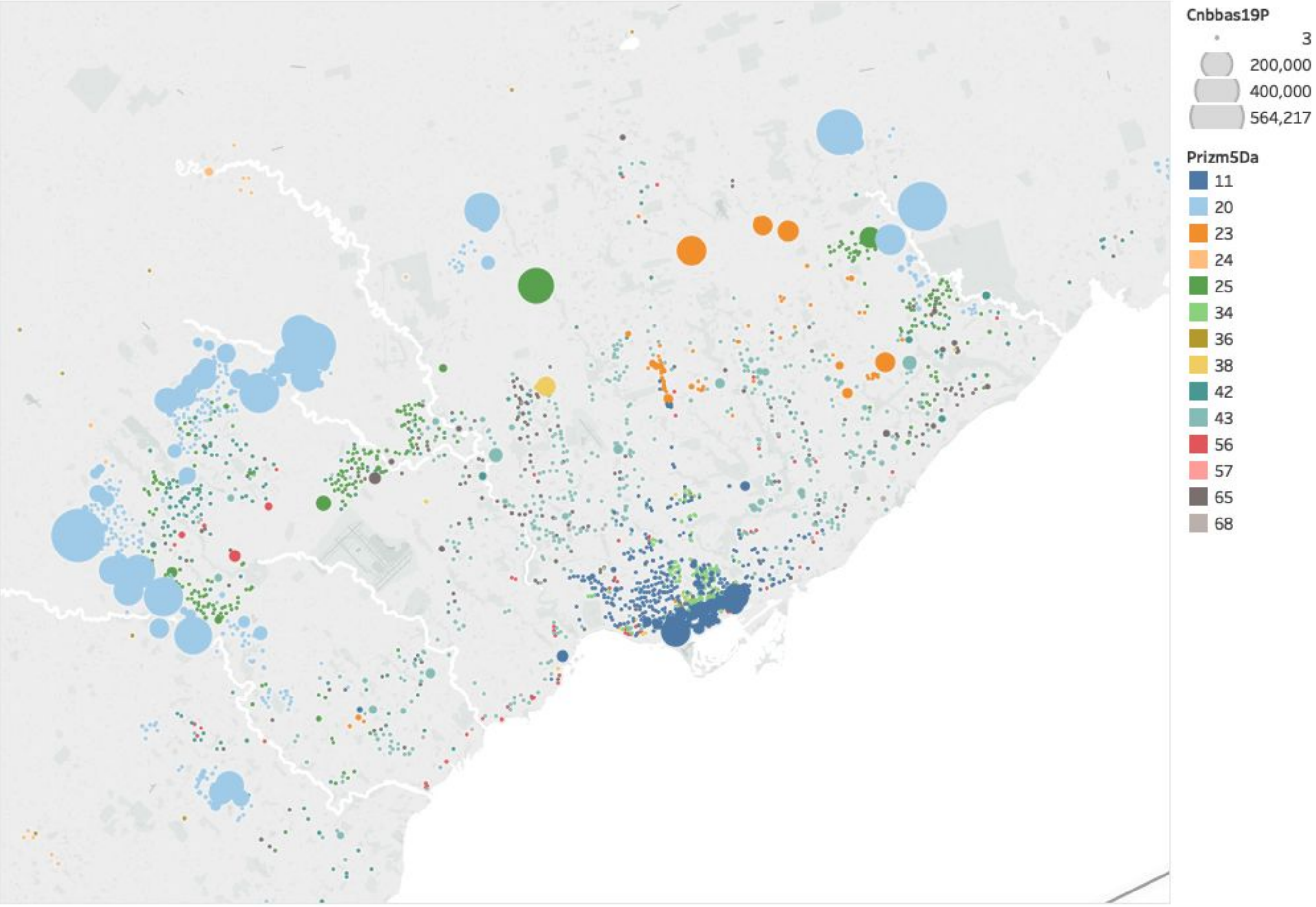
Appendix 3. Example of expected usage per user in DA (Reference for DA 20)



PRIZM5DA	Prediction
20	26.955255
34	19.448445
65	18.083134
38	17.973651
56	17.098997

Map 1

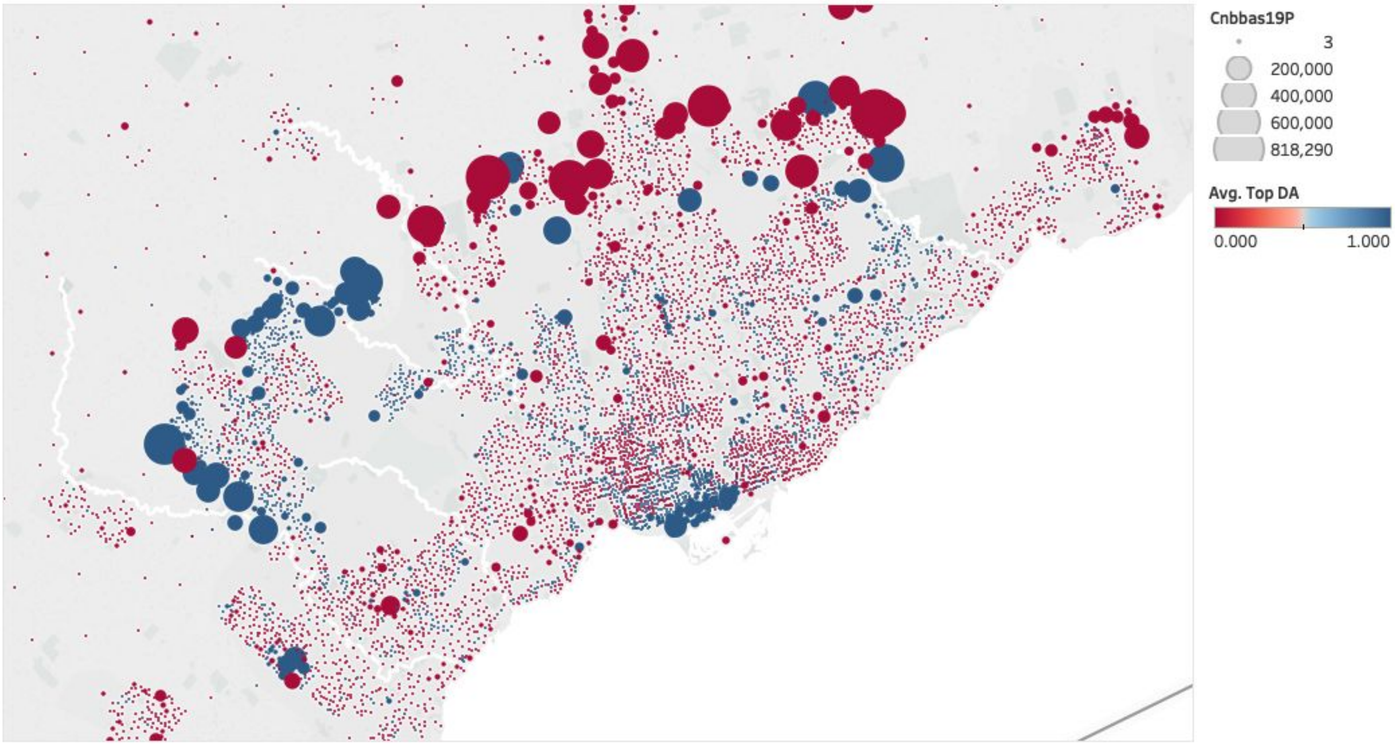
Top DA Distribution



Map based on DArplong/ADlong and DArplat/Adlat. Color shows details about Prizm5Da. Size shows sum of Cnbbas19P. Details are shown for various dimensions. The data is filtered on average of Top DA, which includes values greater than or equal to 1.

Map 2

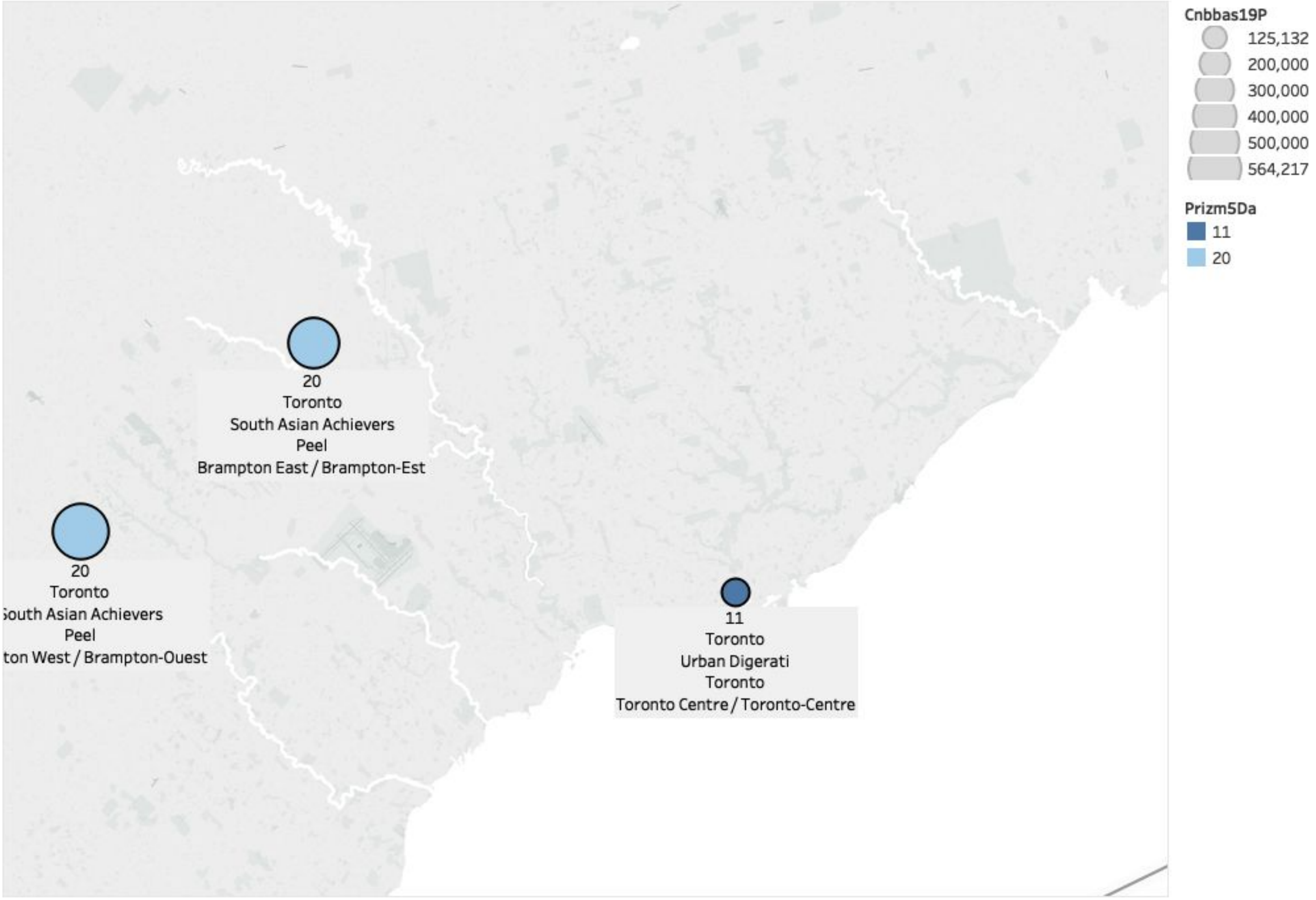
Top vs. Remainder



Map based on DArplong/ADlong and DArplat/Adlat. Color shows average of Top DA. Size shows sum of Cnbbas19P. Details are shown for various dimensions.

Map 3

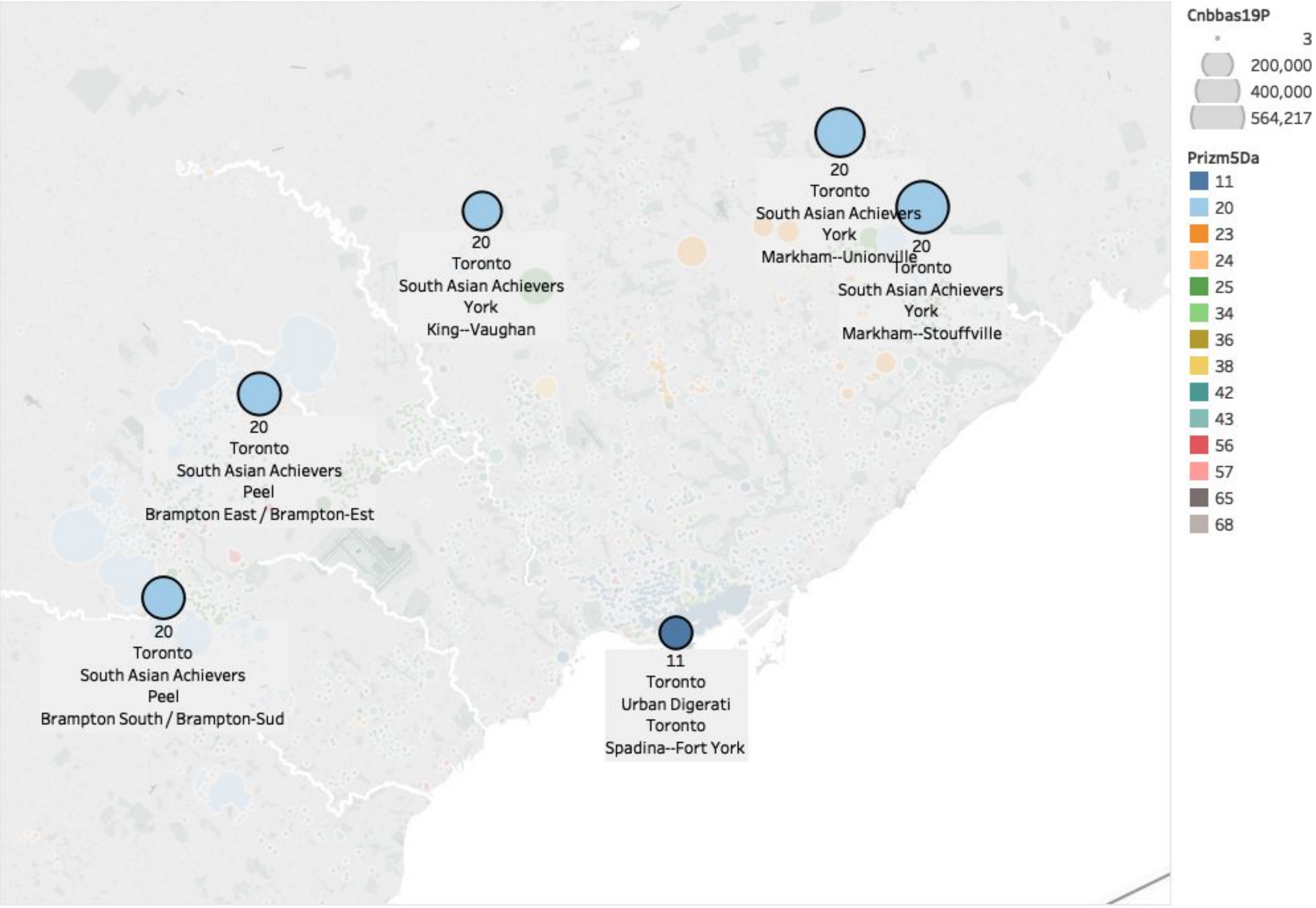
Top DA Distribution



Map based on DArplong/ADlong and DArplat/Adlat. Color shows details about Prizm5Da. Size shows sum of Cnbbas19P. The marks are labeled by Prizm5Da, CMAname/RMRnom, Name, CDname/DRnom and FEDname/CEFnom. The data is filtered on average of Top DA, which includes values greater than or equal to 1. The view is filtered on Inclusions (CDname/DRnom,CMAname/RMRnom,DArplat/Adlat,DArplong/ADlong,FEDname/CEFnom,Name,Prizm5Da), which keeps 3 members.

Map 4

Top DA Distribution



Map based on DArplong/ADlong and DArplat/Adlat. Color shows details about Prizm5Da. Size shows sum of Cnbbas19P. The marks are labeled by Prizm5Da, CMAname/RMRnom, Name, CDname/DRnom and FEDname/CEFnom. The data is filtered on average of Top DA, which includes values greater than or equal to 1.