Aptive

Data Insights & Heatmaps
By ADAlytix

03/06/2019

Agenda

Approach 1

- Data Quality & Summary
- Data Preparation Approach
- Heatmaps & Bubble Charts
- EDA & Insights
- Modeling strategies: Next Steps

Approach 2

- Business Model & Problem Definition
- RFM Analysis
- Charts & Insights
- Customer Churn & Profitability
- Modeling results

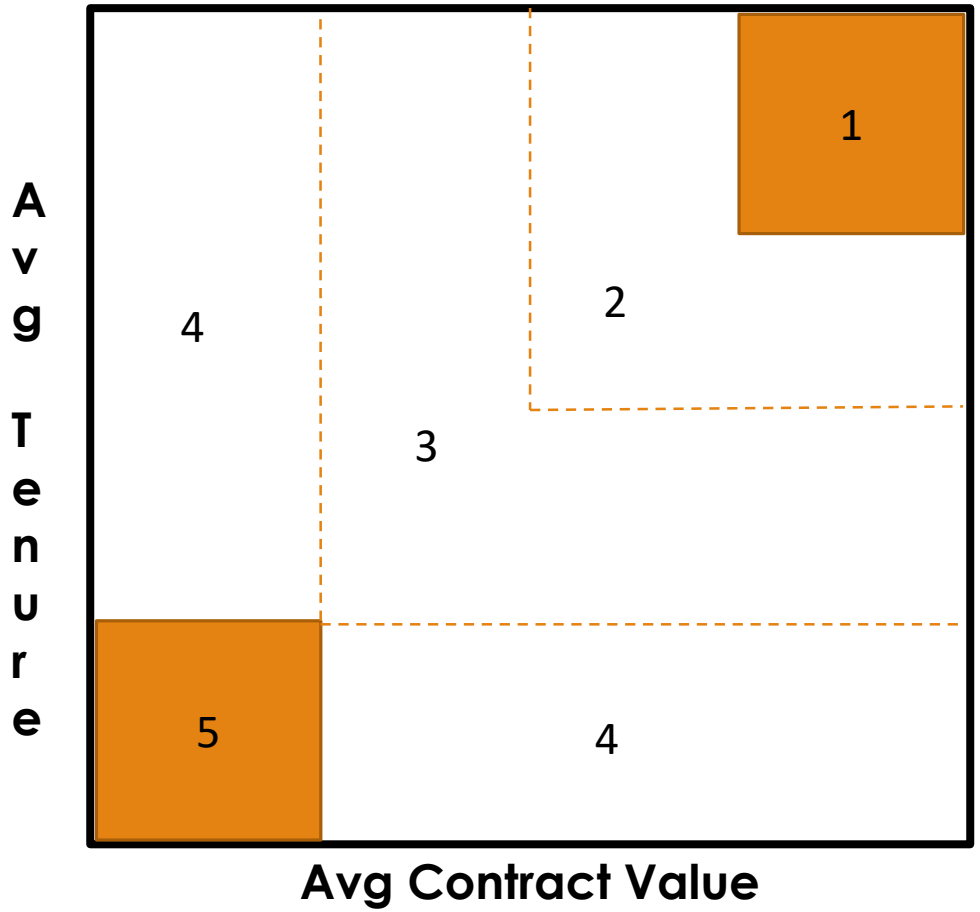
Data Quality & Summary

- **4 cities** – Dallas, Salt Lake city, Washington DC, Houston; **546 zips**; **97,161 customers**
- Demographic variables at zip level; Many customers in a zip
- Avg tenure and Average Contract Value of a customer are of prime importance
- Data Quality issues
  - Some customers had missing demographic data. Imputed with customers from same zip.
  - Some files have zip codes other than the city for which it contains data. For example; 75007,75010 in Carrolton but present in Dallas file. 75052, 75054 present in Grand Prairie but present in Dallas file. 77584 in Pearland but present in

Data Preparation Approach

- Finding zips which has substantial number of customers
- Concatenating all the 4 cities data
- Creating Correct Tenure (in years) for inactive customers based on #services
  *(i/5)*365 if i<= 5 else 365+((i-5)/4)*365; i = services_completed*
- Total Revenue = Avg Contract Value * Correct Tenure
- Removing Nulls with forward fill inside each zip group for missing demographic data
- Interest variables averaged for all customers in zip

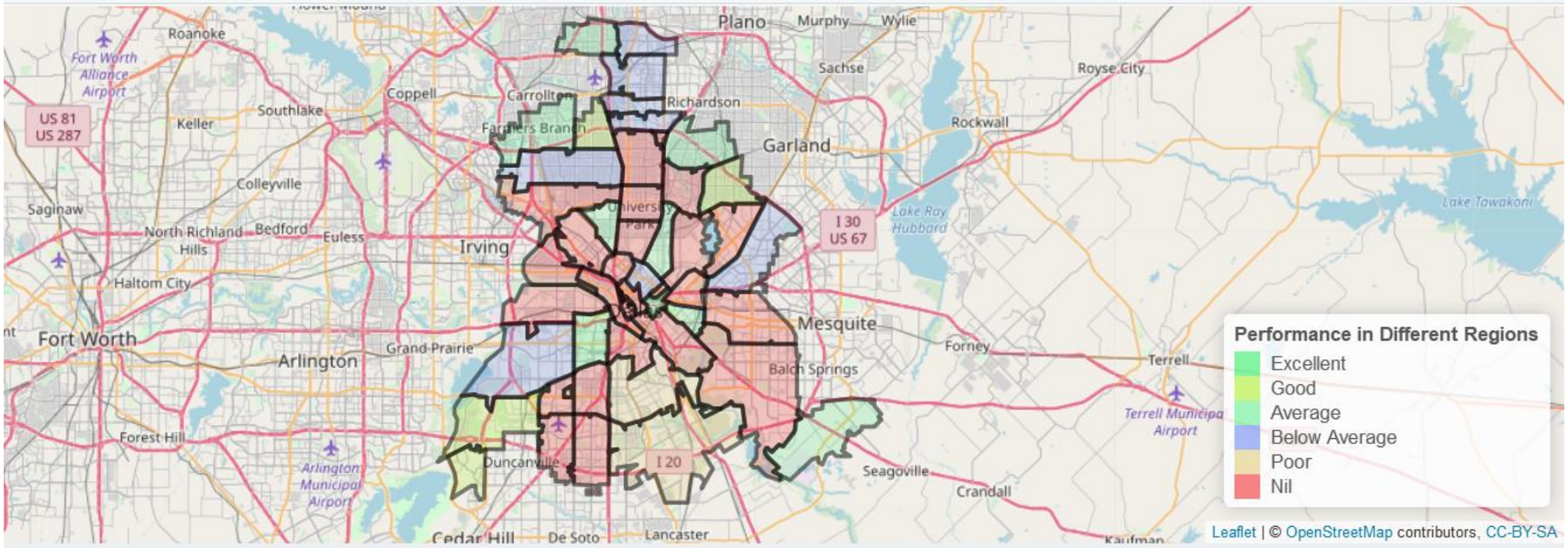
Feature Engineering



1. **Excellent :** P(ACV) > 80 & P(AT) > 80
2. **Good :** 50 < P(ACV) & 50 < P(AT) minus 1
3. **Average:** 20 < P(AT) & 20 < P(AT) minus 1 & 2
4. **Below Average:** 20 < P(ACV) & 20 < P(AT) minus 1,2 & 3
5. **Poor :** P(ACV) < 20 & P(AT) < 20

P – percentile across all the data points
ACV – Avg Contract Value
AT – Average Tenure

Heatmap & Bubble Charts



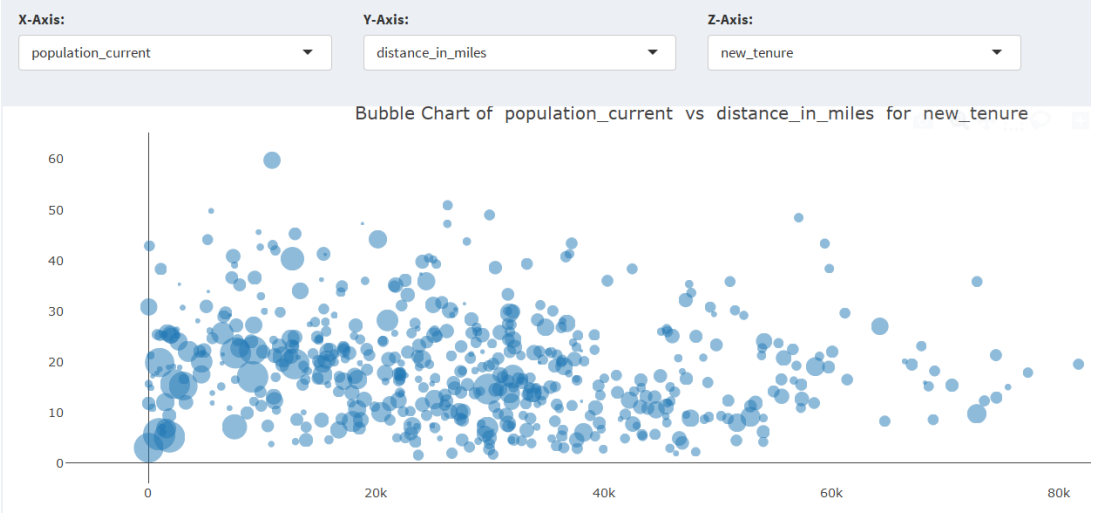
- *Spatial Polygon DataFrames for zip boundaries – available from US gov; used Rgdal library*
- *Leaflet library for interactivity on map*
- *Plotly for scatter and 3D Bubble charts*

Heatmap & Bubble Charts

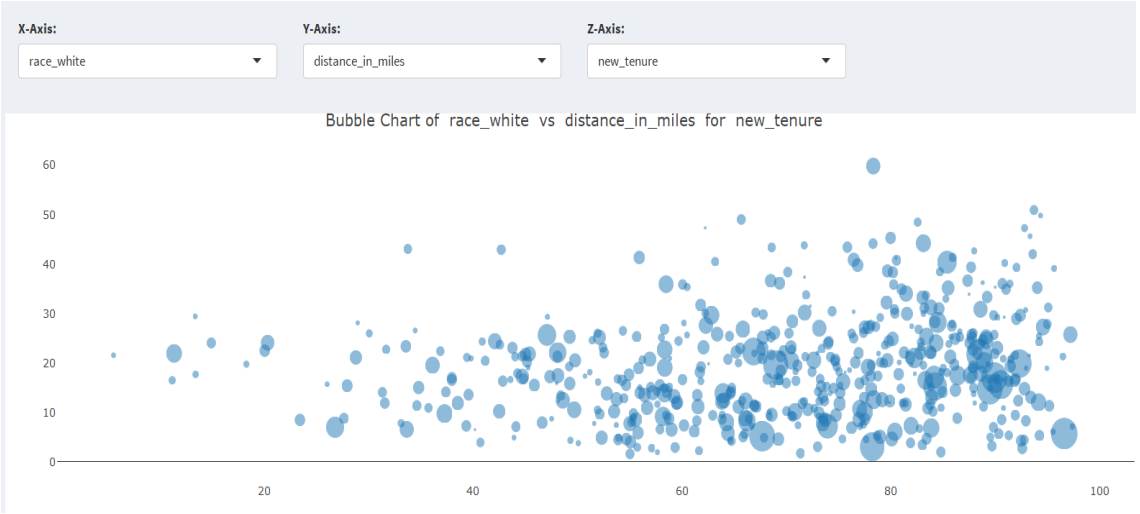
Heatmap & Bubble Charts

Demo

EDA & Insights

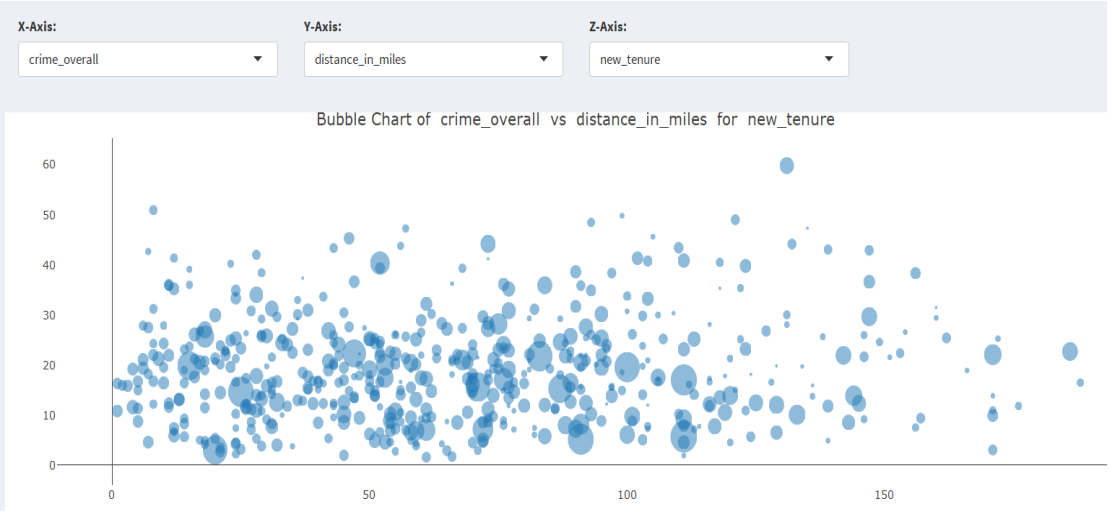


Higher Tenures are concentrated in zips with less population and less distant from office

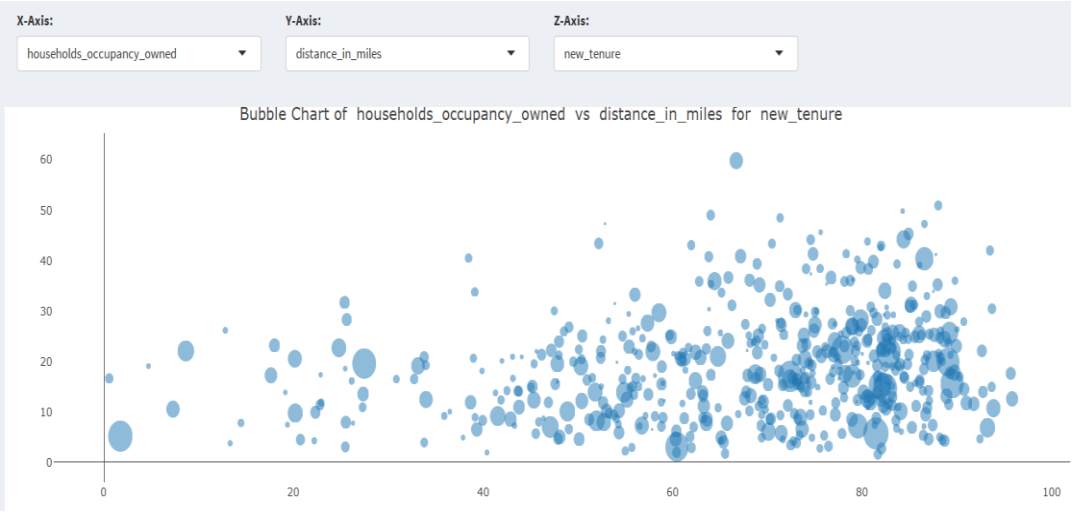


Zips with high % of white people have higher tenures.

EDA & Insights

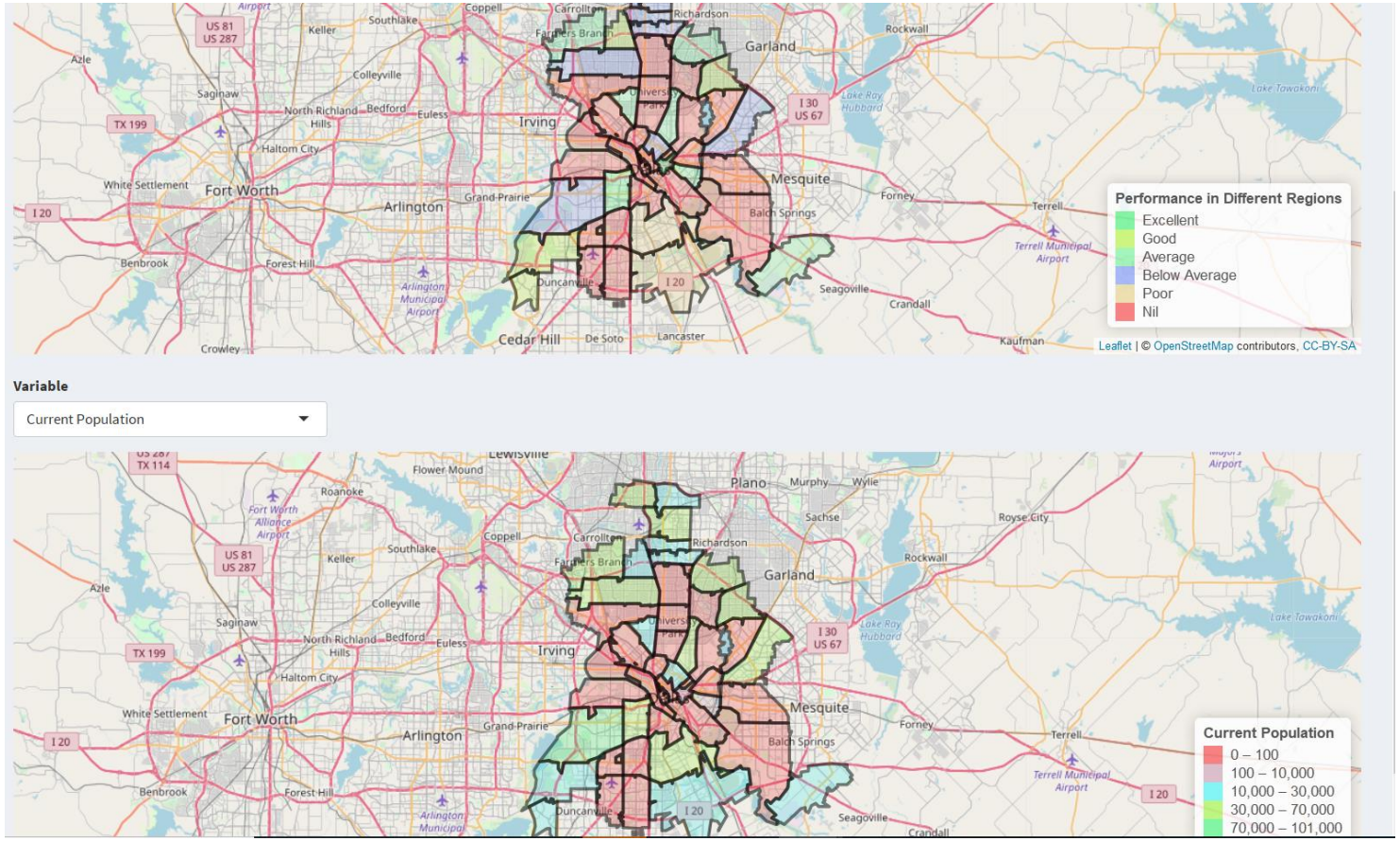


Zips with high crime rates don't have high tenures.



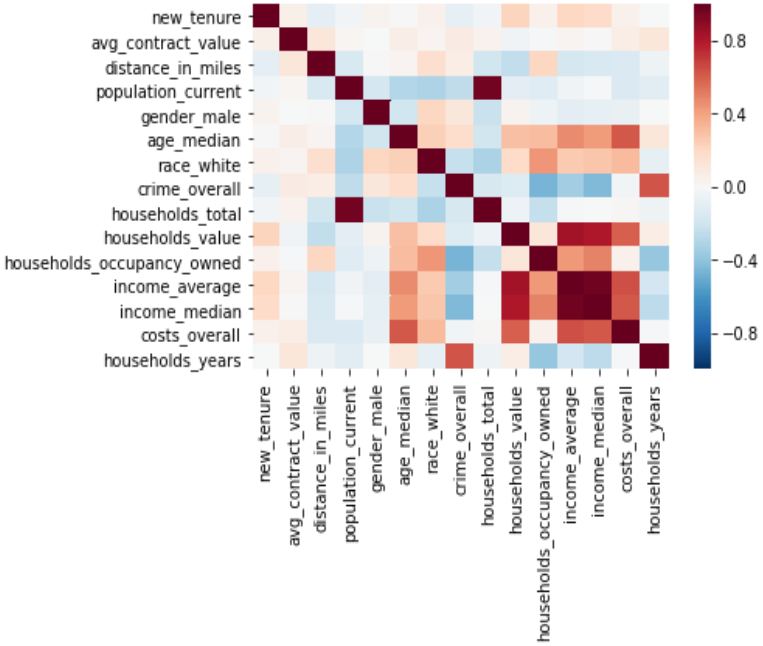
Zips with high house ownership ratios have high tenures.

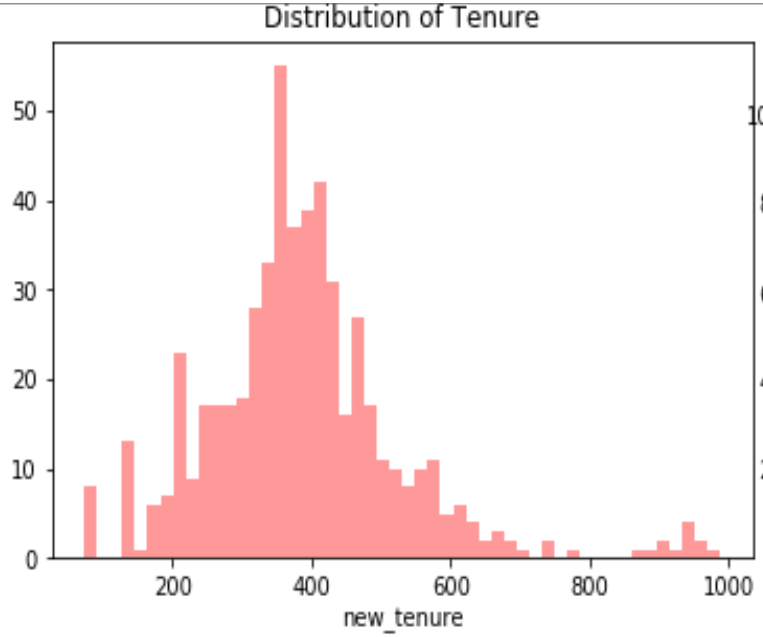
EDA & Insights

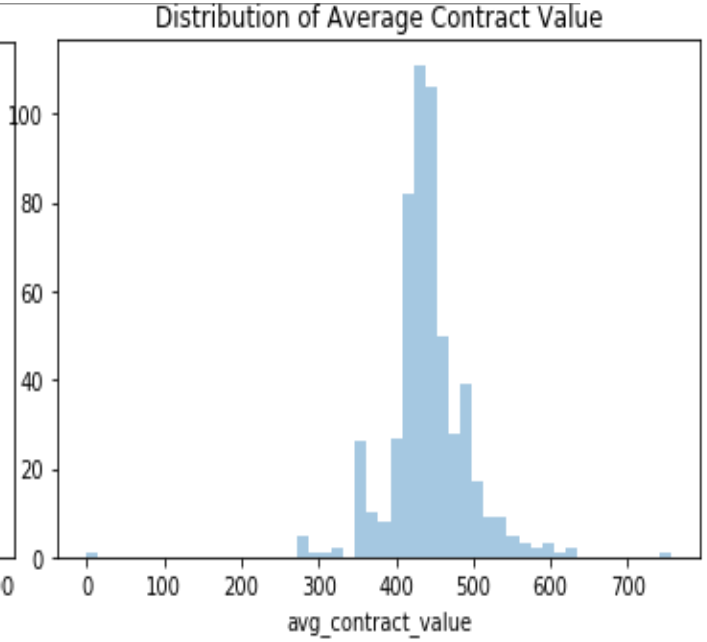


Zips with "Good" performance come from zips with population b/w 10k-30k

EDA & Insights







- "crime_overall" moderate negative correlation with "household_occupancy_owned", "population_current" and "income_average"/"income_median"
- A strong positive correlation exists between "costs_overall" and "households_value", "income_average"/"income_median".
- "age_median" is positively correlated with variables that determine the buying power of customer, like "households_value" and income.

- Mean of tenure = 389, SD of tenure = 143
- Little left skewed; almost symmetric

- Mean of value = 441, SD of value = 53
- Little right skewed

EDA & Insights

	Variables	F_Statistic	p_value
0	new_tenure	14.5347114599	0.0000007079
1	crime_overall	12.3664477776	0.0000055900
2	population_current	5.6462406689	0.0037400918
3	households_years	5.0945631409	0.0064246667
4	households_total	4.2214882120	0.0151586841
5	costs_overall	3.4358628030	0.0328954359
6	households_occupancy_owned	2.2091863178	0.1107704101
7	age_median	2.0298005813	0.1323517902
8	distance_in_miles	1.8193131832	0.1631185743
9	households_value	1.6488688905	0.1932231732
10	income_median	1.1978616264	0.3026309154
11	race_white	1.1260412690	0.3250665502
12	income_average	0.9181695782	0.3998647114
13	gender_male	0.6215877447	0.5374706220

"new_tenure", "crime_overall", "population_current", "households_years", "households_total" and "costs_overall" has impact on "avg contract value"

	Variables	F_Statistic	p_value
0	population_current	11.3816878363	0.0000143634
1	households_total	10.0163542006	0.0000534385
2	crime_overall	6.6705332735	0.0013735696
3	households_value	3.6185019054	0.0274680751
4	households_occupancy_owned	2.7322918391	0.0659598232
5	income_median	2.3790332879	0.0935997856
6	race_white	2.2391001273	0.1075319069
7	income_average	2.2307741740	0.1084236195
8	gender_male	1.5884034633	0.2051937994
9	households_years	1.4122146642	0.2444915439
10	avg_contract_value	1.1302231760	0.3237155617
11	distance_in_miles	0.7984502206	0.4505506092
12	age_median	0.3897657056	0.6774037970
13	costs_overall	0.1494775041	0.8611930387

"population_current", "households_total", "crime_overall" and "households_value" impacts "avg tenure"

Modeling Strategies: Next Steps

- **Regression** (At zip level) – Regression models like Linear/Polynomial Regression, GAM, Regression Trees for predicting Avg Tenure & Avg Contract Value for each zip based on demographic data

- **Customer Survival Analysis** (At customer level) - Finding average tenure for customers using survival analysis model; Cox regression etc.

- **Customer Lifetime Value** – (*Acq. Cost* + *Maint. Cost* – *Total Rev*) for each customers

- Data
  - **Training** – Zip-level aggregated data from 37 cities where Aptive has existing operations
  - **Testing** – Demographic data of the cities where Aptive is considering expanding to. Aggregate at city level to rank cities.

- Cloud – AWS/Azure 16/32 GB RAM Linux VM needed for dashboard deployment

US Pest Control Dataset – Analysis and Recommendations

Problem Definition

As a new business in the Pest Control industry, it is important to identify its most profitable customers and areas with high customer pool vis-à-vis the areas with high customer churn rate and analyse the reasons for the same.

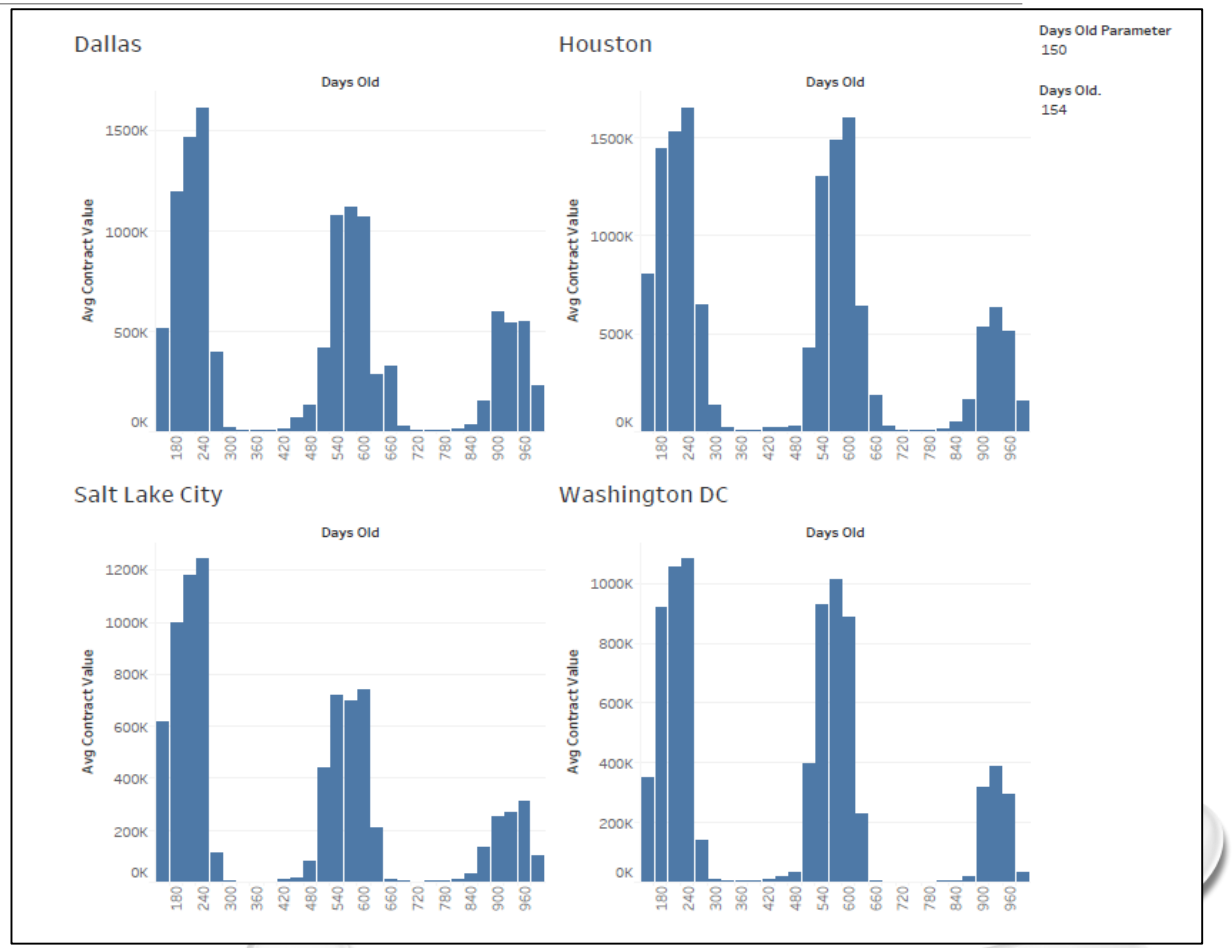
Key areas of concern:

- Identify and target profitable customers and locations based on Customer Lifetime Values and demographics
- estimate the value that you can derive from the customer
- Identify customers who are likely to churn and focus on ways to retain them.

Pest Control Business Model

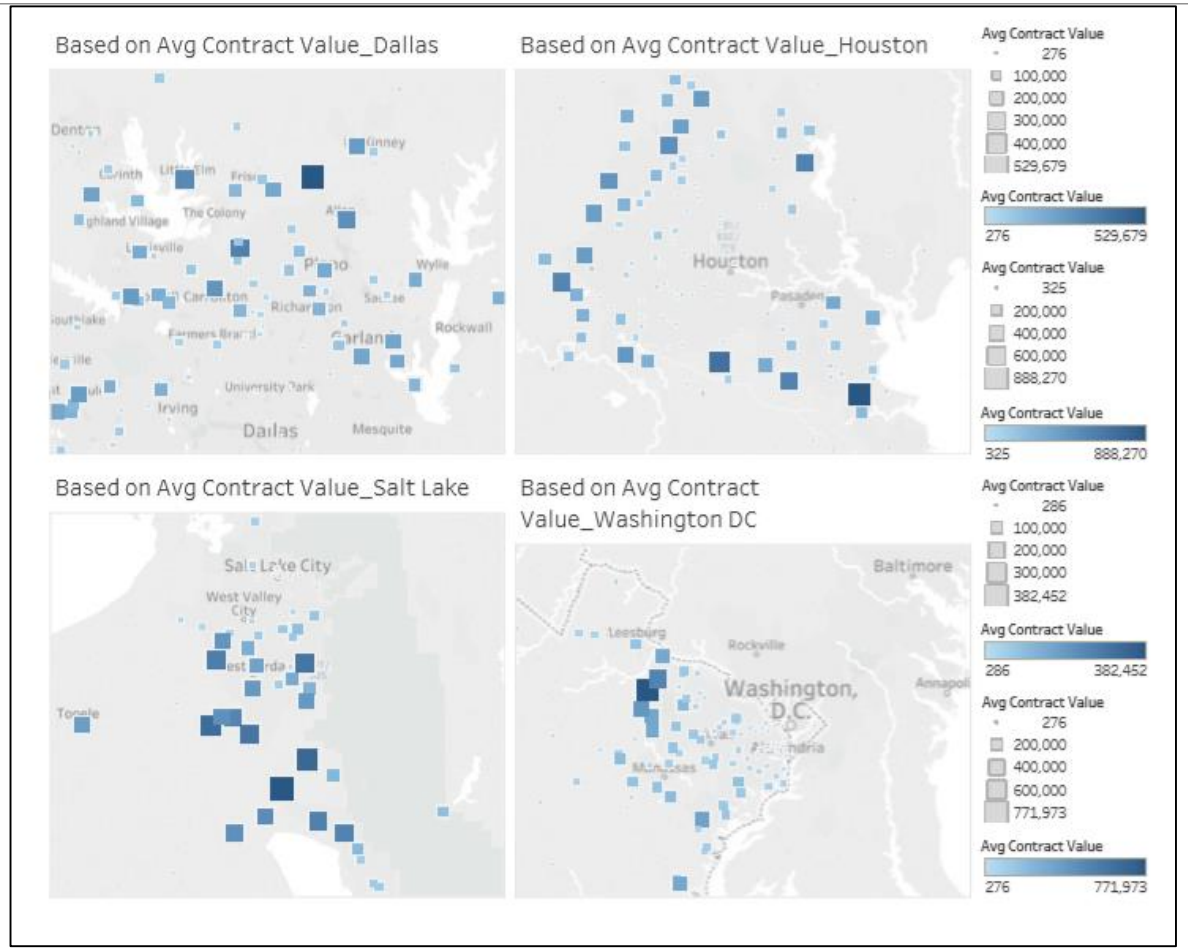
Majority of Door to door sales, cycle peaking during Summer from April-August.

Highest contract values are generated from the most recent acquisitions.



Business Cycle

Geographical Spread – Heat Map Analysis



Heat Map

Recency Frequency Monetary (RFM) Analysis

RFM Analysis

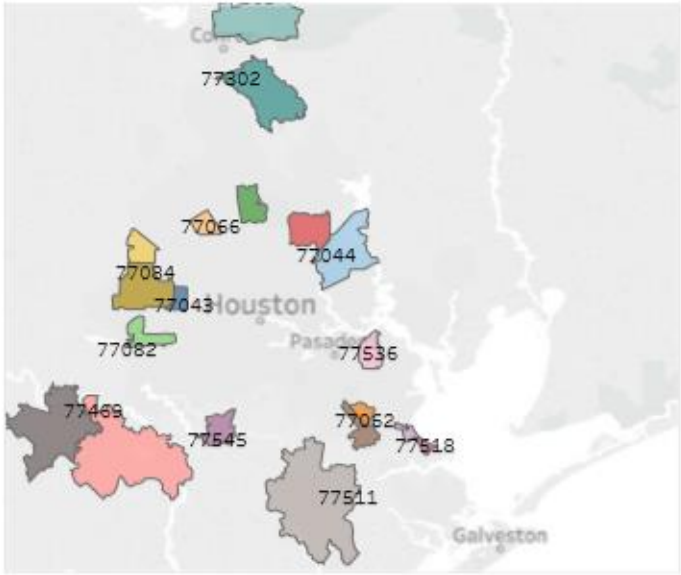
Based on

- Number of Customers (Recency)
- Number of Services Completed (Frequency), and
- Customer Lifetime Value (Monetary),

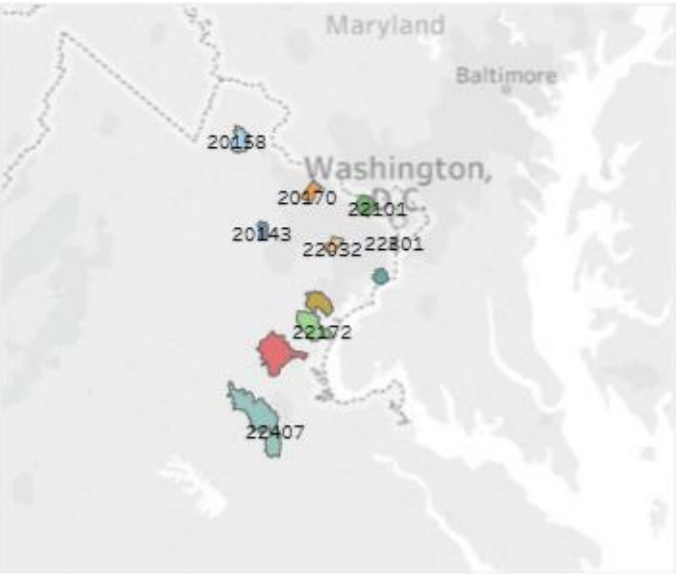
we identified the most profitable localities for the business to focus on

The following heat map indicates top 10% of the areas that contribute the highest to the revenue, thereby enabling the business to narrow down on these households in order to retain them:

Ideal Target Group_Houston



Ideal Target Group_Washington DC

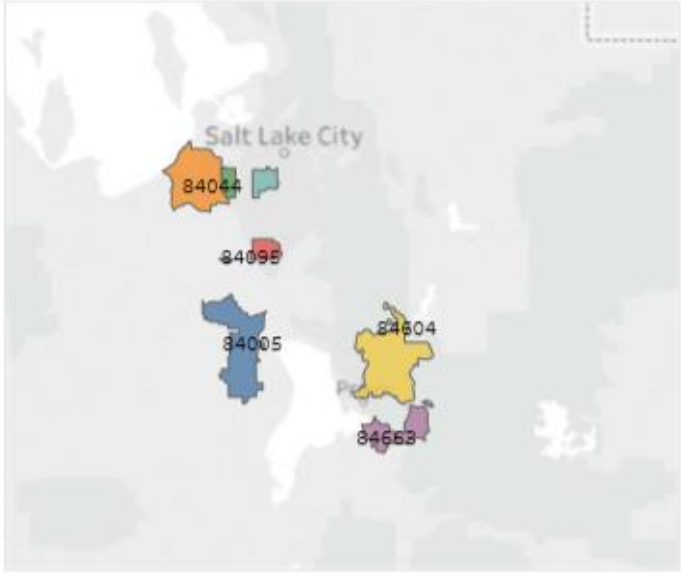


Zip
- 75013
- 75028
- 75040
- 75042
- 75043
- 75044
- 75051
- 75052

Highlight Zip
No items highlighted

Zip
- 77043
- 77044
- 77062
- 77066
- 77073
- 77082
- 77084
- 77095

Ideal Target Group_Salt Lake City



Ideal Target Group_Dallas

Zip
- 84005
- 84044
- 84095
- 84119
- 84128
- 84604
- 84663

Zip
- 20143
- 20158
- 20170
- 22032
- 22101
- 22172
- 22193
- 22301

How to improve Customer Retention

The active customers with a Service Completion Percentage between 0.8 to 1 should be our focus as they have recently stopped availing the services.

Based on the data, major proportion of these customers have an average contract value in the range (300-400) and are in 210-300 days old (3rd Quarter) across all four cities

Therefore the business needs to innovate their strategies in order to retain this proportion of customers

- Introduce invoice due alerts, discounts in advance to the 3rd Quarter to these set of customers.
- Promote the business through cold calling these customers

Retention Analysis (snapshot)

Customers with 0.8-1 completion % in Dallas
(For active customers)

Completion%_Range	0.8-1
Difference Range	(All)

Count of customerID	Column Label										
Row Labels	Q11	Q10	Q9	Q8	Q7	Q6	Q5	Q4	Q3	Q2	Grand Total
1	485	232	14	170	1221	632	57	30	4659	1788	9288
200-300	21	14		5	28	19	3	1	38	21	150
300-400	195	103	2	93	440	266	31	10	1803	981	3924
400-500	85	52	3	46	427	272	16	5	1982	671	3559
500-600	177	63	6	25	280	66	4	12	792	109	1534
600-700	7		2	1	42	6	3	2	41	6	110
700-800			1		4	3			3		11
Grand Total	485	232	14	170	1221	632	57	30	4659	1788	9288

Retention Analysis

Customers to be targeted from the retention analysis (snapshot)

Target Customers List

Target Customers list_Dallas

zip	customerID	days_old	Quarter_Passed	Completion%_Range
75001	987642	252	Q3	0.8-1
75001	983745	253	Q3	0.8-1
75001	957887	263	Q3	0.8-1
75002	1054974	226	Q3	0.8-1
75002	1050091	229	Q3	0.8-1
75002	1068153	221	Q3	0.8-1
75002	1068149	221	Q3	0.8-1
75002	1066193	223	Q3	0.8-1

Churned Customers

Churn Analysis

A prominent pattern of customers who cancel their services lie in the 489-671 days old and have contract values in the range of 300-500 across all four cities

Strategies to prevent churn during this period:

- Introduce promotional offers like discounts and one-time free service to the customers with this kind of attributes well in advance
- Contact the churned customers for feedback and focus on the area of improvement

Customer Churn Zone_Dallas											
(For Inactive Customers)											
Count of customerID	Days old (Quarter wise)										
Average Contract Values	Q11	Q10	Q9	Q8	Q7	Q6	Q5	Q4	Q3	Q2	Grand Tot
0	1414	1302	16	372	3082	2642	79	10	1047	440	10404
>800		1	1		1	1					4
200-300	17	15	1	5	31	26	4		14	15	128
300-400	379	448	5	181	827	826	50	1	401	244	3362
400-500	552	523	1	102	1245	1362	17	4	453	153	4412
500-600	411	295	6	74	821	378	3	5	164	25	2182
600-700	52	18	1	9	149	41	5		14	3	292
700-800	2	2	1	1	8	8			1		23

Model Building - Summary

The classification algorithm yields similar results across all four datasets –

- Algorithm used : Logistic Regression , Random Forest
- Average Model Accuracy : 94.7%
- Most Significant Variables :
  - *completion percentage,*
  - *distance in miles,*
  - *average contract value,*
  - *difference*
- Other significant variables such as *days old*, *median income*, and *households year built* have strong correlation with the target variable and must be taken into consideration while taking business decisions.

Conclusion

The pest control business needs to primarily focus on the low performing regions where customer churn rate is high and work on promoting their services with attractive offers

In regions with high valued customers, the business needs to target the specific set of subscriptions that are most likely to churn and improve on ways to retain them – through promotions, improved reviews, additional services, etc.

As it is a seasonal business, the focus should be on optimizing their resources in order to improve their profit margins. From the above data modelling, an in-depth analysis on the most significant attributes affecting their business can help in identifying the key metrics to improve overall sales.