Census Bureau demographics segmentation for all zip codes

Prepared for Prof. John Sparks

February 23, 2015

UIN:
652518967
665447520
658795799

Executive summary

- This presentation will address the segmentation of Census Bureau Demographics for all zip codes.
- The results are based on data of 33 variables out of 46 variables and 34,297 observations.
- Cluster analysis is used to identify the segments in the data.
- Results show that there are 6 segments of Census Bureau Demographics:
  - Working Well Off
  - Wealthy Segment
  - Young Renters
  - Low Income Non-White Residents
  - Retirement Beneficiaries
  - Rest of America

Summary Figures

- Table 1 below shows the average responses to select census variables for each segment and for the entire set of population

Table 1

Segment	Total Zipcodes	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
Working Well off (1)	7,644	117.62	78.26	21.15	32.95	19.01	46.13	41.34	5.58	83.85	93.04
Wealthy (2)	1,457	157.39	72.38	26.55	37.54	21.91	49.06	32.78	71.23	285.42	88.65
Young Renters (3)	2,047	98.44	56.72	57.87	29.59	19.09	42.71	29.62	12.39	104.66	80.15
Low Income Non-White Residents (4)	2,892	70.37	75.13	34.13	39.54	24.35	48.92	44.89	3.78	57.09	39.57
Retirement Beneficiaries (5)	9,040	75.91	71.35	23.07	49.66	33.63	54.53	31.89	1.80	45.33	94.35
Rest of America (6)	10,297	75.86	77.40	22.94	36.61	22.59	47.13	41.82	2.28	55.17	94.57
Overall	**33,377**	**89.91**	**74.27**	**25.83**	**39.17**	**24.67**	**48.87**	**38.14**	**6.67**	**72.33**	**88.25**

- Table 2 provides a summary measure of the segment averages relative to the overall average.
- For Example, Average Income Index was 117.62 for the Working Well off segment vs. 89.91 for the entire population. This translates into a relative measure of 31% ((117.62-89.91)/89.91 = 0.31 or 31%).

Table 2

Segment	% of Zipcodes per cluster	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
Working Well off (1)	23%	31%	5%	-18%	-16%	-23%	-6%	8%	-16%	16%	5%
Wealthy (2)	4%	75%	-3%	3%	-4%	-11%	0%	-14%	968%	295%	0%
Young Renters (3)	6%	9%	-24%	124%	-24%	-23%	-13%	-22%	86%	45%	-9%
Low Income Non-White Residents (4)	9%	-22%	1%	32%	1%	-1%	0%	18%	-43%	-21%	-55%
Retirement Beneficiaries (5)	27%	-16%	-4%	-11%	27%	36%	12%	-16%	-73%	-37%	7%
Rest of America (6)	31%	-16%	4%	-11%	-7%	-8%	-4%	10%	-66%	-24%	7%

Segment 1 – Working Well Off

- The figures for the First segment are shown below.
- The relative percentage of “Income Index” is 31% higher than the average figure across all zip codes.
- The relative percentage of “People 65 Plus” is 23% lower than the average figure across all zip codes.
- Based on the above points, we can say that this segment can be referred as “Working Well Off”.

Segment	% of Zipcodes per cluster	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
1	23%	31%	5%	-18%	-16%	-23%	-6%	8%	-16%	16%	5%

Segment 2 – Wealthy

- The figures for the Second segment are shown below.
- The relative percentage of “Household Valued 200k Plus” is 968% higher than the average figure across all zip codes.
- The relative percentage of “Owner Occupied Median Home Value” is 295% higher than the average figure across all zip codes.
- Based on the above points, we can say that this segment can be referred as “Wealthy”.

Segment	% of Zipcodes per cluster	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
2	4%	75%	-3%	3%	-4%	-11%	0%	-14%	968%	295%	0%

Segment 3 – Young Renters

- The figures for the Third segment are shown below.
- The relative percentage of “Households That Are Families” is 24% lower than the average figure across all zip codes.
- The relative percentage of “Rent” is 124% higher than the average figure across all zip codes.
- The relative percentage of “People 55 to 64” is 24% lower than the average figure across all zip codes.
- The relative percentage of “Median Age” is 13% lower than the average figure across all zip codes and value is lowest among all the segments.
- Based on the above points, we can say that this segment can be referred as “Young Renters”.

Segment	% of Zipcodes per cluster	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
3	6%	9%	-24%	124%	-24%	-23%	-13%	-22%	86%	45%	-9%

Segment 4 – Low Income Non-White Residents

- The figures for the Fourth segment are shown below.
- The relative percentage of “White Residents” is 55% lower than the average figure across all zip codes.
- The relative percentage of “Income Index” is 22% lower than the average figure across all zip codes and value is lowest among all segments.
- Based on the above points, we can say that this segment can be referred as “Low Income Non-White Residents”.

Segment	% of Zipcodes per cluster	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
4	9%	-22%	1%	32%	1%	-1%	0%	18%	-43%	-21%	-55%

Segment 5 – Retirement beneficiaries

- The figures for the Fifth segment are shown below.
- The relative percentage of “People 55 to 64” is 27% higher than the average figure across all zip codes.
- The relative percentage of “People 65 Plus” is 36% higher than the average figure across all zip codes and value is highest among all segments.
- The relative percentage of “Median Age” is 12% higher than the average figure across all zip codes and value is highest among all segments.
- The relative percentage of “Income index” is 16% lower than the average figure across all zip codes.
- Based on the above points, we can say that this segment can be referred as “Retirement beneficiaries”.

Segment	% of Zipcodes per cluster	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
5	27%	-16%	-4%	-11%	27%	36%	12%	-16%	-73%	-37%	7%

Segment 6 – Rest of America

- The figures for the Sixth segment are shown below.
- The relative percentage of “Zip codes per cluster” is 31% higher than the average figure across all zip codes.
- This segment has highest number of zip codes present compared to the remaining segments.
- The relative percentage for the remaining variables are not deviating much from the Overall average values.
- Based on the above points, we can say that this segment can be referred as “Rest of America”.

Segment	% of Zipcodes per cluster	Income Index	Households That Are Families	Rent	People 55 to 64	People 65 Plus	Median Age	Household with a person Under 18	Household Valued 200K Plus	Owner Occupied Median Home Value	White Residents
6	31%	-16%	4%	-11%	-7%	-8%	-4%	10%	-66%	-24%	7%

Summary

- Results of Cluster Analysis based on 33 variables showed 6 segments in the census bureau demographics
    - Working Well Off
    - Wealthy Segment
    - Young Renters
    - Low Income Non-White Residents
    - Retirement Beneficiaries
    - Rest of America
- Technical details regarding the cluster analysis are contained in the Appendix.

Technical Appendix

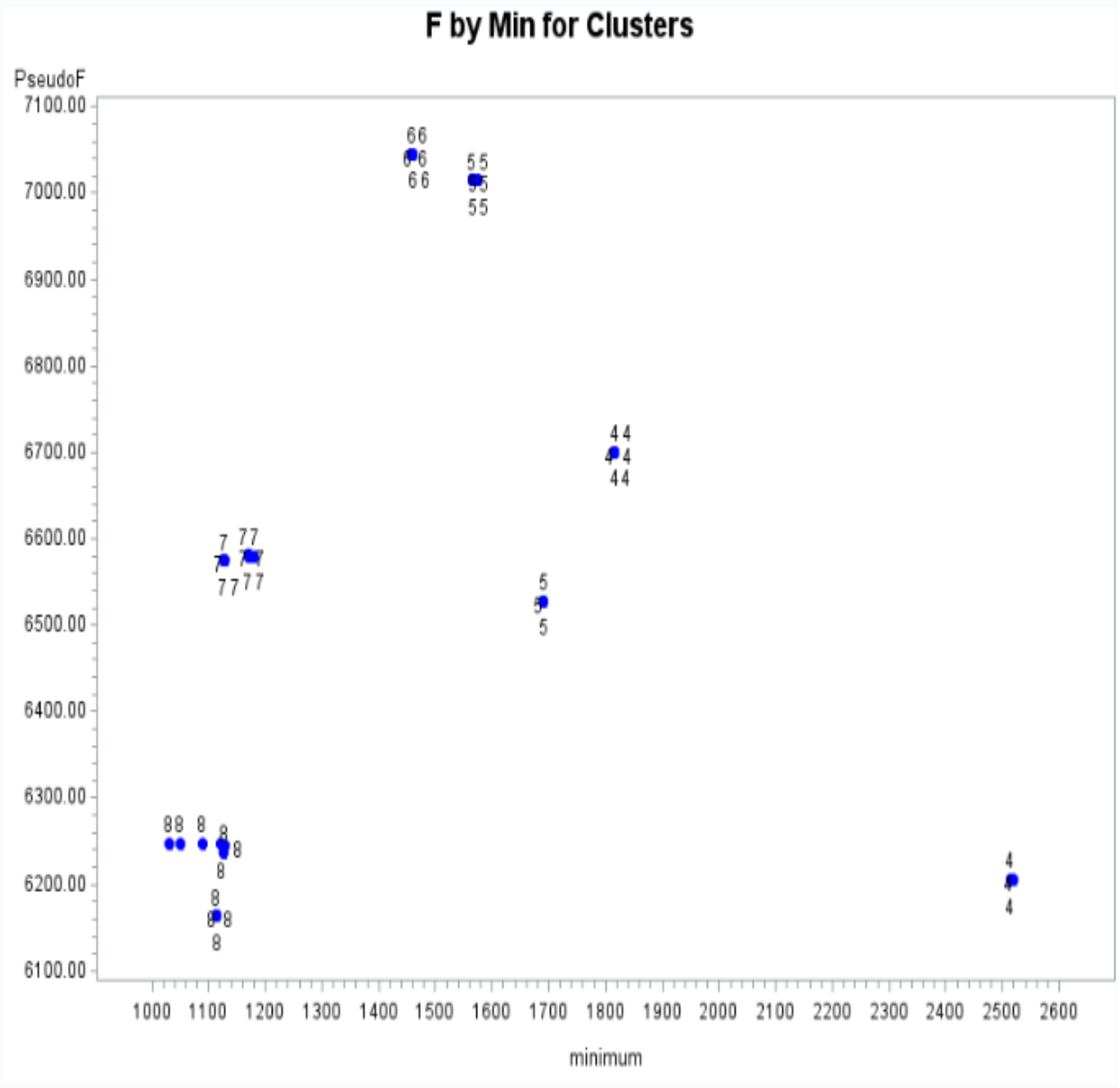
Methods

- The previous analysis is executed using K-means cluster analysis.
- Hierarchical cluster analysis was not used because the number of observations were more in number (34,297) and it will be used only if observations are less in number (smaller data sets).
- The inputs to the cluster analysis are factor scores.
- These were chosen rather than variables because the factor scores from a factor analysis are orthogonal, or uncorrelated.

Multiple Iterations

- The K-means algorithm was executed 10 times for number of clusters 4 through 8.
- Each iteration had a different random starting point.
- Two important criteria for selecting the correct number of clusters are Discrimination and Stability because:
  - Large pseudo F-statistic implies that the segments have close knit distribution and are farther apart from each other. This also means a higher discrimination.
  - A minimal change in clusters which is measured by change in pseudo F-statistic value and size of small segment tells us about the stability of the segment.

Pseudo-F By Size of Smallest Segment



- The graph at left shows the Pseudo F-statistic and Size of Smallest Segment for each of the 10 iterations for 4 through 8 segments.
- We eliminate the 7 and 8 segment solutions because the pseudo F-statistic value is decreased for cluster 7 and cluster 8 as the number of clusters increases from 6. This implies that for cluster 7 and cluster 8 doesn't have good discrimination.
- We can also eliminate the 4 segment solution because of inferior discrimination and inferior stability when compared with cluster 5 and 6.
  - Average of pseudo F-statistic for cluster 4 is less than Average of pseudo F-statistic for cluster 5.
  - Stability based on pseudo F-statistic is also less for cluster 4 solution than cluster 5 solution because cluster 5 solution is grouped closer together when compared to cluster 4 solution

Detailed Analysis of the 5 and 6 segment solution

Minimum	Pseudo F	Iter	Clusters
1,691	6,527.45	501	5
1,564	7,015.02	502	5
1,691	6,527.45	503	5
1,566	7,015.02	504	5
1,575	7,015.01	505	5
1,564	7,014.96	506	5
1,564	7,015.02	507	5
1,564	7,015.02	508	5
1,564	7,015.02	509	5
1,691	6,527.45	510	5
1,459	7,045.33	601	6
1,460	7,045.33	602	6
1,460	7,045.33	603	6
1,457	7,045.33	604	6
1,457	7,045.33	605	6
1,460	7,045.33	606	6
1,460	7,045.33	607	6
1,457	7,045.33	608	6
1,460	7,045.33	609	6
1,457	7,045.33	610	6

Pseudo F-statistic comparison:

- Average pseudo F-statistic value for cluster 5 is 6,527.45
- Average pseudo F-statistic value for cluster 6 is 7,045.33
- Since the average pseudo F-statistic value is higher for cluster 6, we can say that cluster 6 has superior discrimination.

Detailed Analysis of the 5 and 6 segment solution

Minimum	Pseudo F	Iter	Clusters
1,691	6,527.45	501	5
1,564	7,015.02	502	5
1,691	6,527.45	503	5
1,566	7,015.02	504	5
1,575	7,015.01	505	5
1,564	7,014.96	506	5
1,564	7,015.02	507	5
1,564	7,015.02	508	5
1,564	7,015.02	509	5
1,691	6,527.45	510	5
1,459	7,045.33	601	6
1,460	7,045.33	602	6
1,460	7,045.33	603	6
1,457	7,045.33	604	6
1,457	7,045.33	605	6
1,460	7,045.33	606	6
1,460	7,045.33	607	6
1,457	7,045.33	608	6
1,460	7,045.33	609	6
1,457	7,045.33	610	6

Stability based on pseudo F-statistic comparison:

- From the table for cluster 5 solution, we can see that for 3 iterations pseudo F-statistic value is around 6,527.45 and for 7 iterations it is around 7,015.02
- For all the 10 iterations in cluster 6 solution, the pseudo F-statistic value is same.
- As the solution for all different iterations resulted in the same ending point for cluster 6, we can say that cluster 6 has superior stability based on pseudo F-statistic.

Detailed Analysis of the 5 and 6 segment solution

Minimum	Pseudo F	Iter	Clusters
1,691	6,527.45	501	5
1,564	7,015.02	502	5
1,691	6,527.45	503	5
1,566	7,015.02	504	5
1,575	7,015.01	505	5
1,564	7,014.96	506	5
1,564	7,015.02	507	5
1,564	7,015.02	508	5
1,564	7,015.02	509	5
1,691	6,527.45	510	5
1,459	7,045.33	601	6
1,460	7,045.33	602	6
1,460	7,045.33	603	6
1,457	7,045.33	604	6
1,457	7,045.33	605	6
1,460	7,045.33	606	6
1,460	7,045.33	607	6
1,457	7,045.33	608	6
1,460	7,045.33	609	6
1,457	7,045.33	610	6

Stability based on Size of smallest segment comparison:

- For cluster 5 solution, largest value of size of smallest segment is 1,691 and smallest value of size of smallest segment is 1,564. Therefore maximum difference in the number of zip codes in the smallest cluster is 127. This implies that from one iteration to other, 127 zip codes are changing membership in and out of smallest cluster.
- For cluster 6 solution, largest value of size of smallest segment is 1,460 and smallest value of size of smallest segment is 1,457. Therefore maximum difference in the number of zip codes in the smallest cluster is 3. This implies that from one iteration to other, only 3 zip codes are changing membership in and out of smallest cluster.
- Since less number of zip codes are changing membership for cluster 6 solution when compared to the cluster 5 solution, we can say that cluster 6 has the superior stability based on size of smallest segment.

Team

- Mounica Sirineni – 652518967
- Sai Dheeraj Illendula – 665447520
- Sai Lahari Jalaparthi – 658795799