Data Challenge 2020: Analytics Process & Trends from a review of HUD Programs, 2009-2018

|  |  |  |  |
| --- | --- | --- | --- |
| **DaCruz Norberto, MS**  University of Maryland  R.H. Smith School of Business  College Park, MD  dacruz.norberto@rhsmith.umd.edu | **Daniel Zhu**  University of Maryland  Computer Science  College Park, MD  dzhu@terpmail.umd.edu | **Richard Zhou**  University of Maryland  Computer Science  College Park, MD  rzhou564@terpmail.umd.edu | **Elizabeth Bonsignore, Ph.d**  University of Maryland  Human-Computer Interaction Lab  College Park, MD  ebonsign@umd.edu |

# ABSTRACT

This paper presents the analysis, process, and findings from Data Challenge 2020, a multi-institution Data Analytics “Datathon'' hosted by the University of Maryland’s College of Information Studies (“Maryland’s iSchool”). The results below uncover changes over time, patterns, anomalies, and insights that team 200047 discovered while performing various data analyses on a subset of non-pii HUD program data between 2009-2018. The data contains household-level demographic data for HUD’s largest housing assistance programs. The data includes household characteristics of household, income, and program variables enabling researchers to perform analyses not possible using other datasets.

This article describes data exploration, analytic methodologies, model development, results and detailed list of lifecycle recommendations. This data analysis will compare rent burdens across demographics and programs.

The outcome of these analyses will consider HUD’s four strategic goals and be associated to the sustainability of providing more affordable, discrimination-free, quality of living for all groups of people. The team’s primary end-goal aims to determine data and policy recommendation for HUD regarding mitigate any gaps in the composition of public-housing assistance recipients in the program. In addition, the team will use predictive analytic techniques to forecast if:

1. Program vouchers will be able to defray the cost of rising housing costs continuing to cover 60% of costs with linear regression
2. Use Decision Analysis to determine uncover business rules and patterns in those receiving public housing
3. Predict the composition of applicants/recipients for each program type in the next five years

## Author Keywords

Public Housing, Vouchers, Visualization, R, Machine Learning, Clustering, Decision Tree, Data Analytics, Analysis, Housing and Urban Development, Federal Subsidy Programs, Government, Federal.

# INTRODUCTION

As denoted in HUD’s 2018-2022 four-year strategic plan, “HUD is working to strengthen the housing market to bolster the economy and protect consumers; meet the need for quality affordable rental homes; utilize housing as a platform for improving quality of life; build inclusive and sustainable communities free from discrimination,” [1]. HUD’s strategic goals employ the principles of social and economic sustainability. HUD supports the modification of communities to promote sustainable living across the price and cost dimensions of economic sustainability. HUD accomplishes this by assisting recipients though HUD’s largest housing assistance program categories including: Housing Choice Vouchers (HCV), Multifamily programs (MF), and Public Housing (PH). Strategically, HUD is interested in determining trends regarding household characteristics for those participating in these programs (e.g., combined household incomes, composition of children/adults, racial and ethnic backgrounds, etc). The main goal of the team’s analysis aims to support the transformation of assisted housing by transitioning 125,000 Public Housing units to a more sustainable platform from FY18 through the end of FY20 [1].HUD’s overarching research question for Data Challenge teams analyzing its dataset:

* How have the household characteristics (e.g., household composition, income, etc.) of HUD’s PH, MH, and HCV programs changed from 2009 to 2018?

Team 20047 extrapolated several sub-questions from this initial framing is tied to HUD’s strategic initiatives, and asking:

* Can we predict what sorts of households will either need more assistance or no longer be able to afford their rent, even with assistance?
* What are differences and trends (if any) regarding the HCV program,

## The HUD Dataset

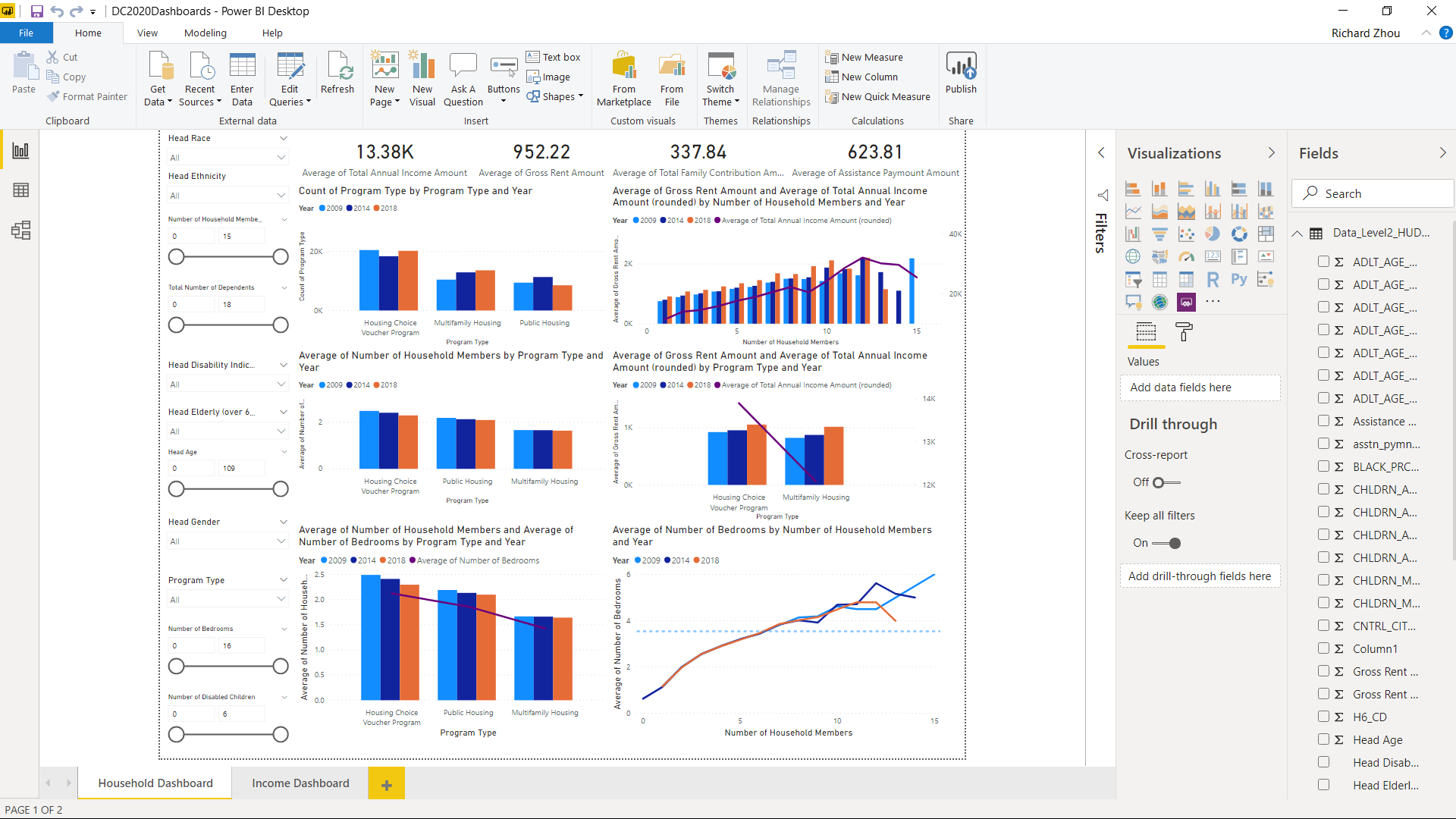
The HUD provided three datasets that each contained a random sample of more than 40,000 households in HUD’s Public Housing (PH), Multifamily Housing (MH), and Housing Choice Voucher (HCV) programs in 2009, 2014 and 2018, respectively. A data dictionary was also provided that included detailed descriptions of the column headings for each file. Overall, the HUD administrative files contain housing, income, and program participation data for recipients of MF, HCV, and PH programs in all states, the District of Columbia, and some territories (for example, Puerto Rico and the U.S. Virgin Islands).

## DATA EXPLORATION and CLEANSING

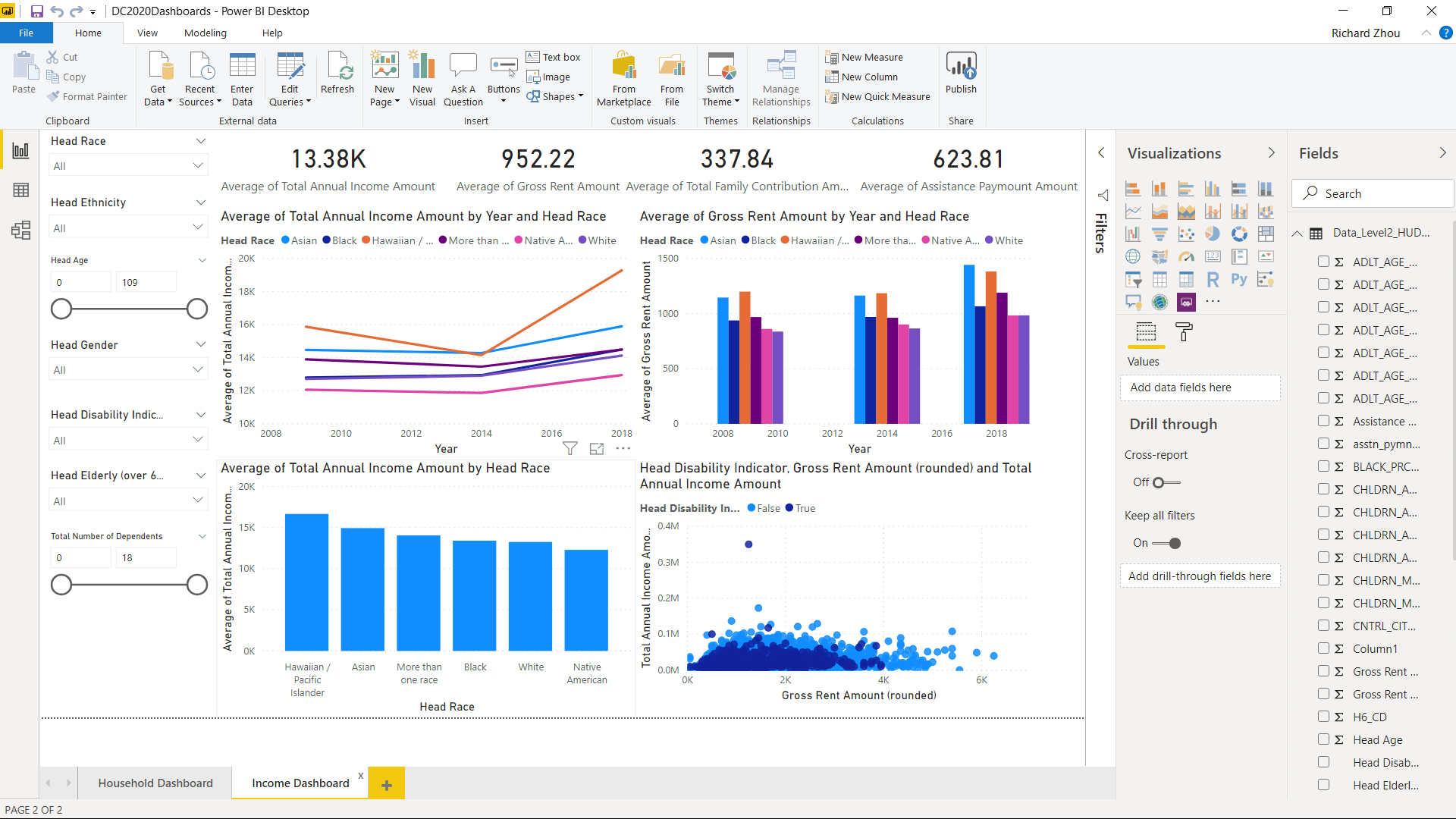
During initial analysis, the team discovered several irregularities in the dataset that required cleaning and removal. Essentially, a large amount of rental information regarding the program information, such as “gross rent” and HUD’s assistance payments were not entered correctly and were represented by a “.” (columns: gross\_rent\_amnt, gross\_rent\_amnt\_rounded, HUD’s asstn\_pymnt\_amnt\_ rounded). The team determined that these data required cleaning; however, one of the concerns was that removing these data points might affect or skew the overall analysis in some way. The team’s solution was to create two data sets, one with the rows that had erroneous data removed (i.e., a clean set) and one that only contained rows that had the “.”, or the erroneous data. If the demographic information for each was similar, then the removal would effectively **not** skew the analysis, it would just remove errors and missing data. Overall, troughly 50% of the data was unusable fromhe original set.

The team used PowerBI in order to create dashboards to conduct early data exploration. The Houseboard Dashboard compares various household variables and the Income Dashboard examines annual income and gross rent to other variables. The team found that there was a significantly higher number of White and Black people under any of the three housing plans (representing 52% and 43% of the data respectively). The number of children (people under the age of 18) has dropped across all 3 years, the number of adults from 18 to 50 have slightly decreased, while those older than that have slightly increased: most notably, there has been a significant increase of adults of age 62 to 85. Overall, there was no change or a very slight decrease in the total number of people in the households. Female heads tended to have a higher number of total dependents. Number of bedrooms and total number of household members were strongly correlated. People with disabilities tended to have lower annual income, gross rent, and average assistance payment amount. Both income and rent have increased between 2009 to 2018 regardless of race. There was also a significantly higher number of people under the Housing Choice Voucher Program followed by Multifamily Housing then Public Housing. A notable exception are Asians who have a higher number of families in the Multifamily Housing program. Public Housing did not have data on Average Assistance Payment Amount.

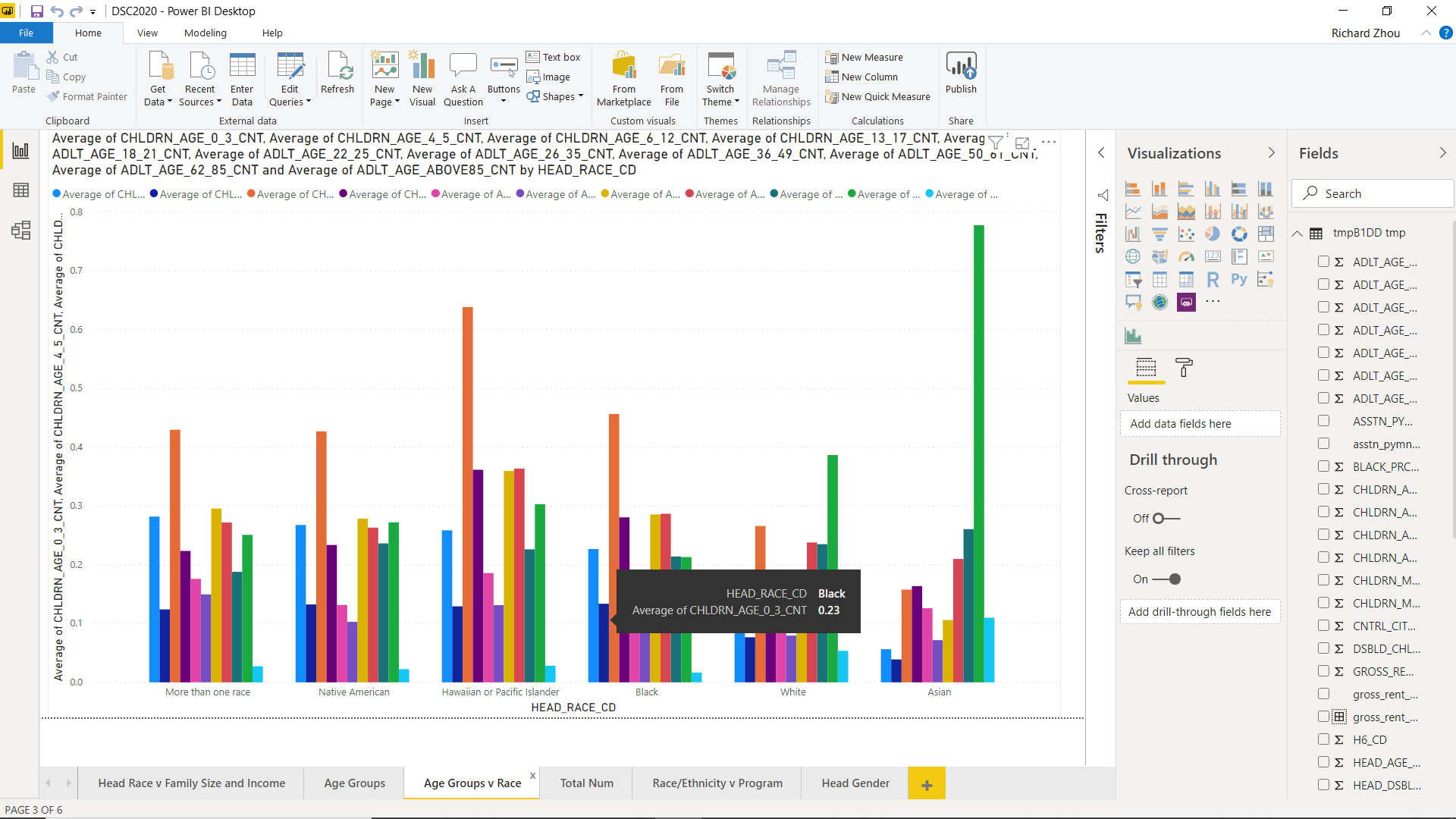
*Household Dashboard*

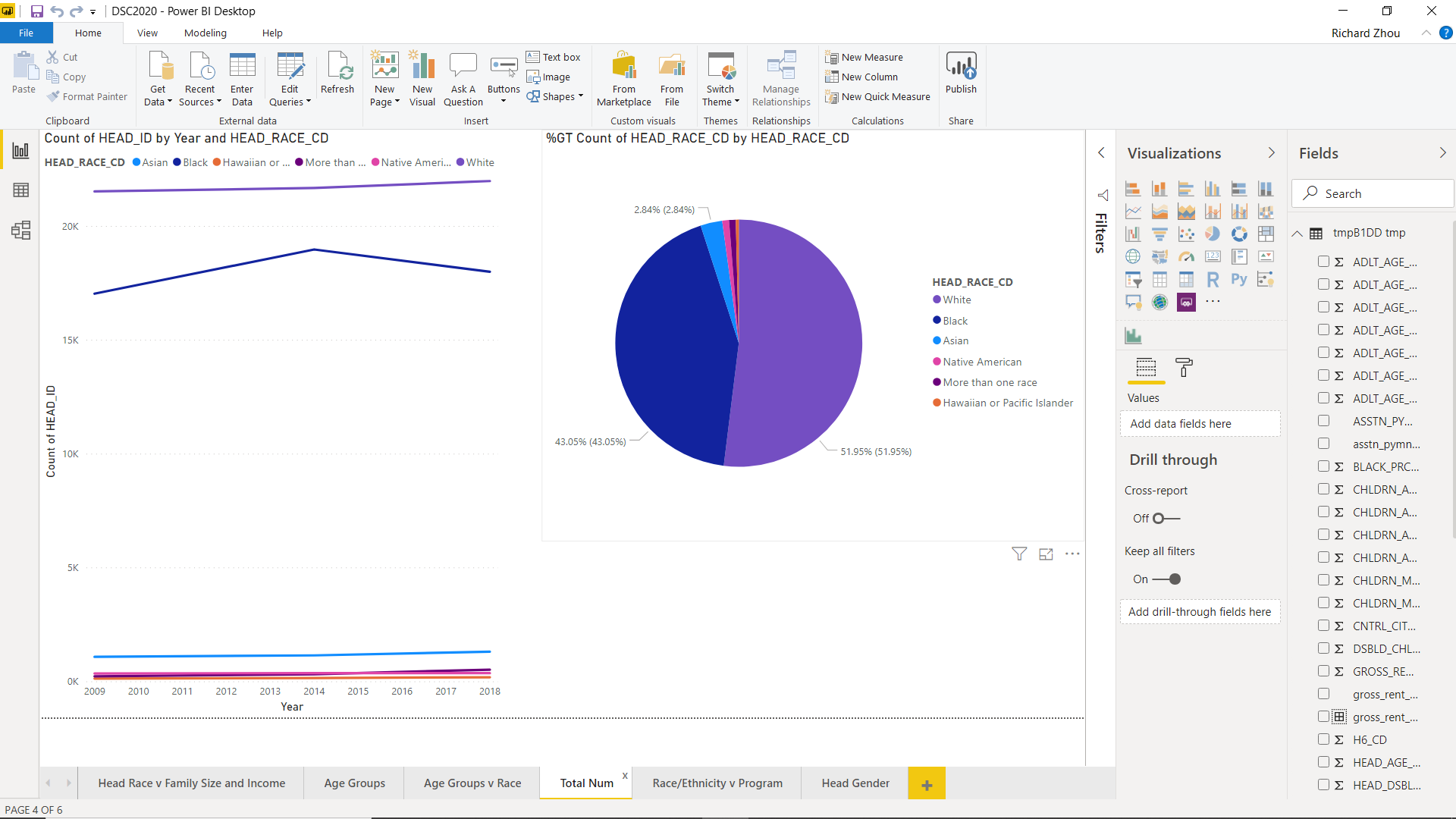
**

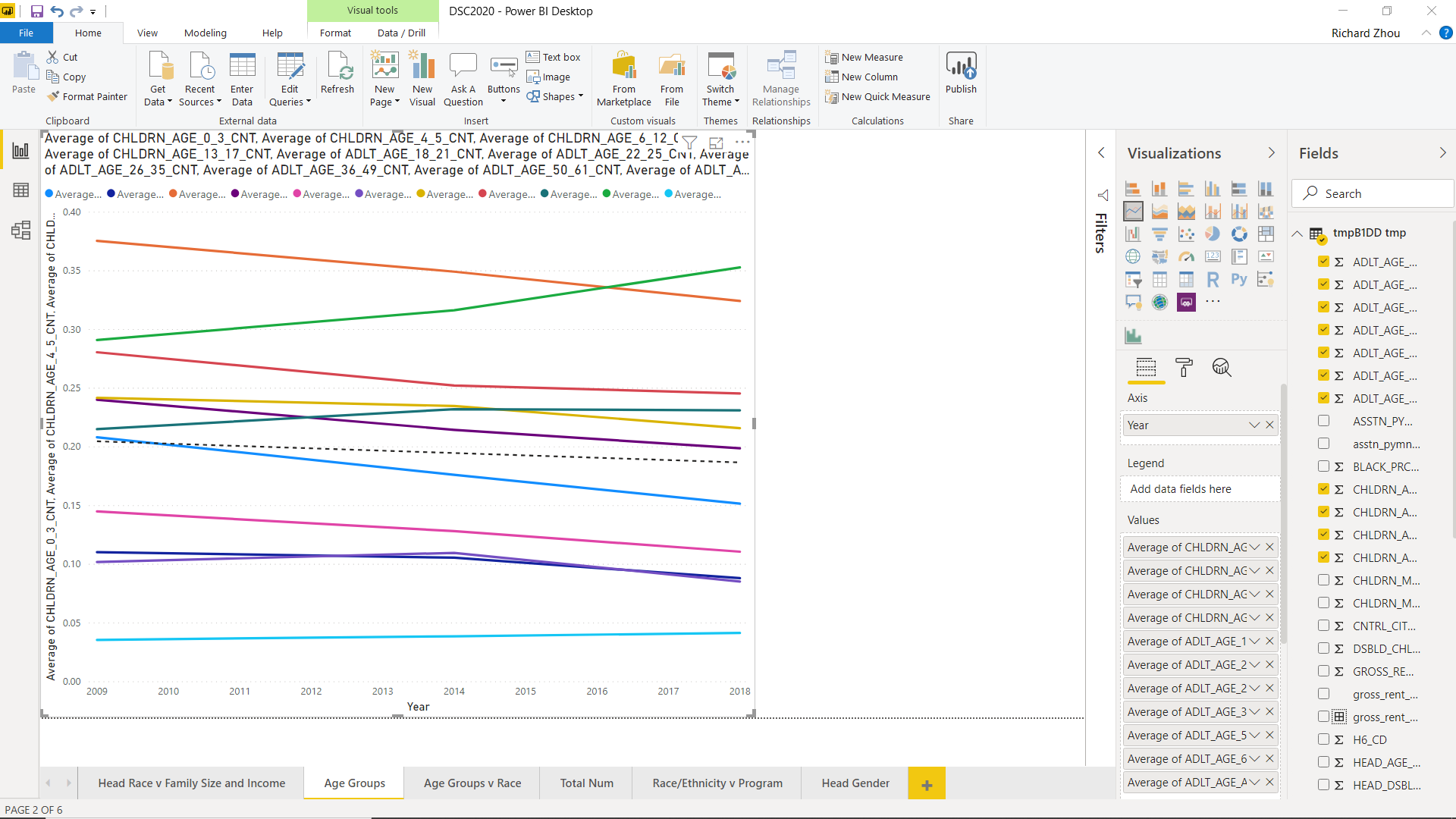
*Income Dashboard*

**

*Additional Figures*







## Household Demographic Trends

* Household sizes have decreased across all races overtime
* The number of children (people under the age of 18) has dropped across all 3 years, the number of adults from 18 to 50 have slightly decreased, while those older than that have slightly increased: most notably, there has been a significant increase of adults of age 62 to 85. Overall, there was no change or a very slight decrease in the total number of people in the households.
* Native Americans, Hawaiian and Pacific Islander, Blacks, and people of more than one race have more children in the housing plan (especially of ages 6 to 12) while Whites and Asians have more people in the 62 to 85 age range. This is especially notable for Asians.
* The average incomes are ranked from greatest to least as follows:
  + Hawaiian and Pacific Islander, Asian, More than One race, Black, White, Native American (This is the same order as the gross rent).

## Program Trends

* Rent data was only collected for the Housing Choice Voucher Program and Multi-Family Housing but not Public Housing
* Gross Rent has increased for all races over time
* The number of families with housing plans have increased for all groups across all 3 years. Number of black families increased from 2009 to 2014 but decreased in 2014 to 2018, but it’s still an overall increase from 2009 to 2014.
* Incomes levels in Housing plan 3 (Multifamily housing) were noticeably lower than plans 1 (Public Housing) and 2 (Housing Choice Voucher Program) - these 2 had about the same average income levels

## IncomeTrends

* There’s a sudden drop in income for Hawaiian/Pacific Islanders in 2014 (remains about the same between 2009 and 2018 overall) - (why did this happen?)
* Asian, Black, and White incomes increased between all 3 years, Native American incomes increased from 2009 to 2014 and remained stagnant between 2014 to 2018, mixed race income has remained about constant/slightly decreases

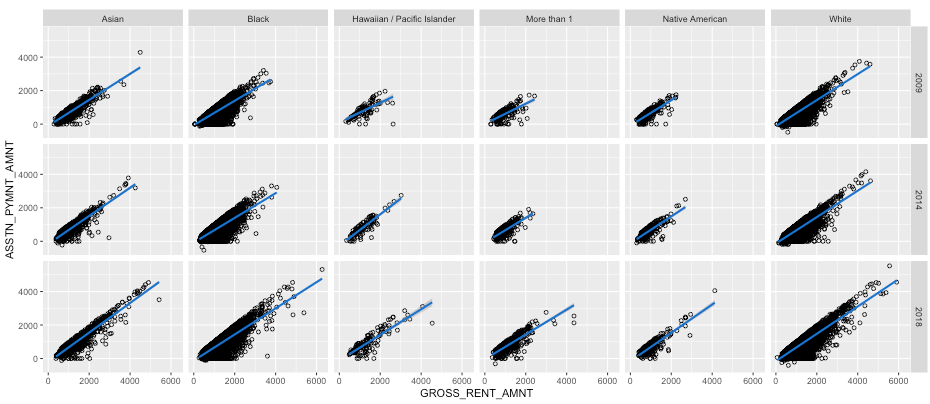
# METHODS AND MODEL DEVELOPMENT

The team conducted a series of machine learning analysis using the RStudio Program

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Learning  Method | Description | Problem Type | Strengths | Weaknesses |
| Decision Trees | Hierarchical learning that splits data in branches to maximize information gain at each split. | Regression or Classification | Can learn non-linear relationships  Scalable and handle outliers well  Can model non-linear boundaries hierarchically | Unconstrained  Prone to overfitting |
| Linear Regression | Predict the value of an outcome variable Y based on one or more input predictor variables X | Regression | Straightforward, easy to understand and explain, and can normalized to avoid overfitting | Limited to linear relationships and looks at the mean of dependent variables. Sensitive to outliers and multicollinearity |
| K-means Cluster Analysis | Organizing disorganized objects based on properties | Classification | Provides an unbiased look how data can be grouped | Not as sophisticated as other techniques and assume all clusters are equal in size |

**Decision Tree Analysis:** The Decision Tree breaks down various demographic details in order to best predict a household’s program.

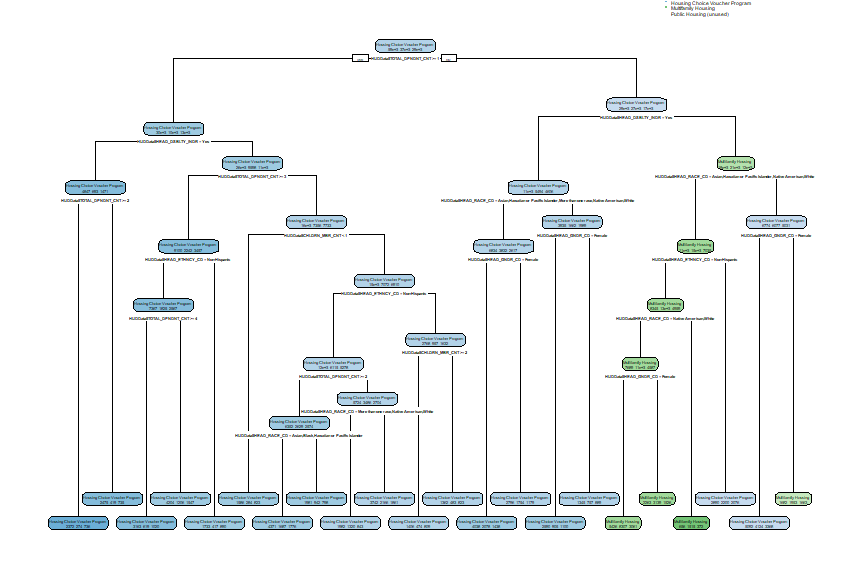
**Linear Regression:** Linear regression is the modeling of data using a single independent and dependent variable, which is used to predict future data points. Linear regression provides a simple yet important method of understanding the relationship between two variables. For this project, two linear regression models were generated: one with the independent variable represented by the field “year” while the field “asstn\_pymnt\_amnt” represented the dependent variable, and one with the independent variable “gross\_rent\_amnt” and the dependent variable “asstn\_pymnt\_amnt. The “asstn\_pymnt\_amnt” field represents the amount of money HUD pays for the rent, which differs from household-to-household, and the “gross\_rent\_amnt” represents the amount of money households were charged for their rent. In addition, the data was further separated based on the race of the head of the household, which is divided into “White”, “Black”, “Native American”, “Asian”, “Hawaiian / Pacific Islander”, and “More than 1”. The aim of this is to see how much assistance would different ethnic groups receive in the future and how that may be impactful towards HUD’s subsidized housing programs.



**Cluster Analysis:** Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Clustering provides a systemic approach to grouping variables and data elements based on similar and dissimilar characteristics. In order to improve data quality and ensure proper development of the cluster analysis model, the data preprocessing phase to develop a wine quality clustering analysis included using a ‘scale’ method available in the ‘cluster’ package in r was used on the dataset so that all variables are being compared with the same data type and standard. The field ‘program\_type\_edited’ is removed since it’s a classifying field for recipients in their respective program types. To ensure that the k-means cluster algorithm runs correctly, all fields are checked for a value that is not null. The objective of the unsupervised learning approach is for the algorithm to generate its own classification based on similarities and dissimilarities within the data. “The aim is also to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis,” [2]. Figure 2 of the Appendix shows the dataset pre-scaled and the data post-scaled in respective .csv files.

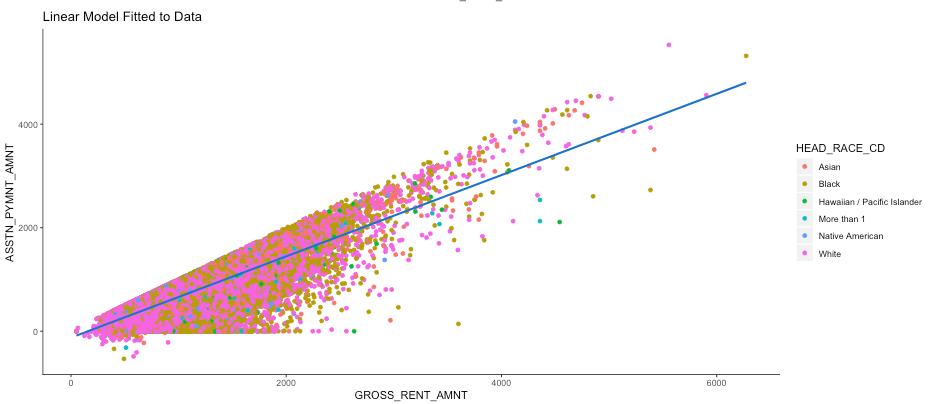
# RESULTS OF MODEL ANALYSIS

**Decision Trees:**



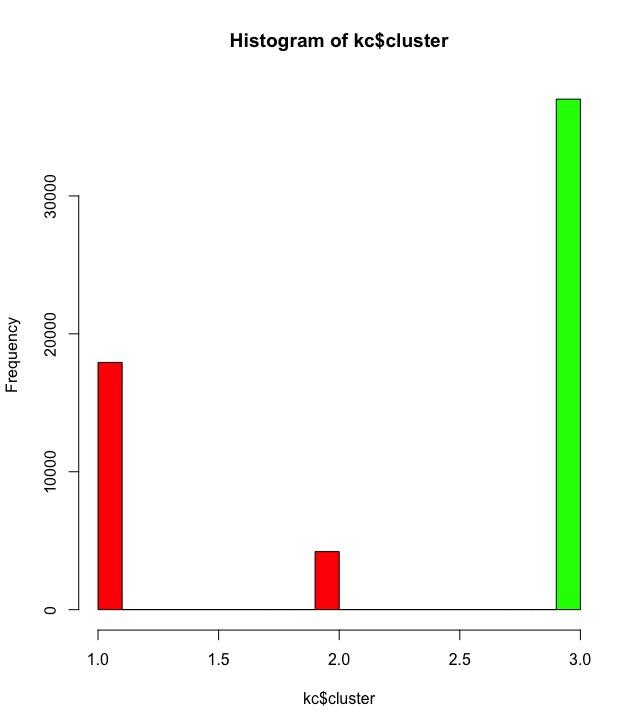
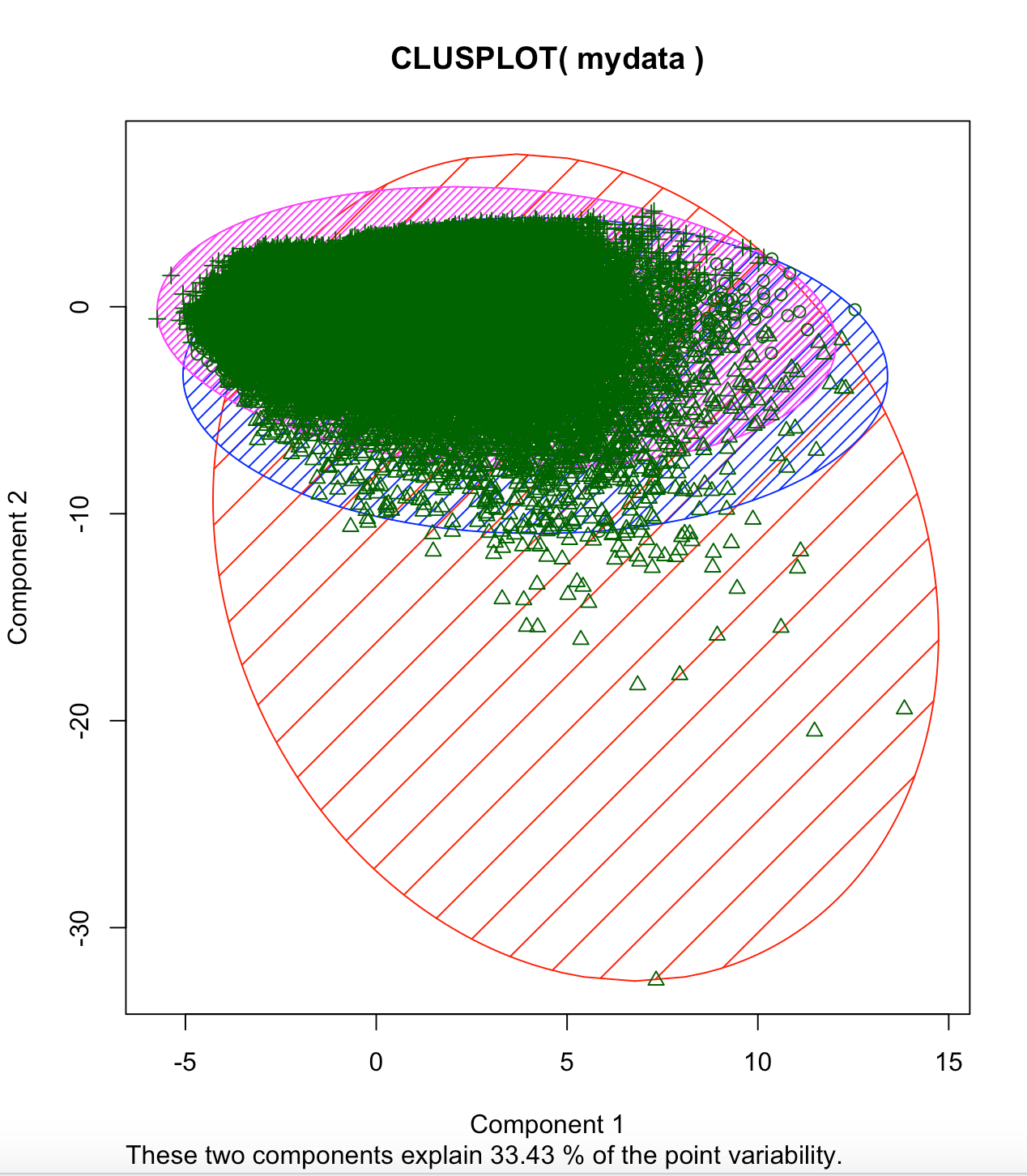
The decision tree only broke down into either the Housing Choice Voucher Program or Multifamily Housing program. Earlier data exploration indicated that these two programs had the most number of people enrolled, and the Decision Tree seems to suggest that the available factors do not allow us to distinguish the Public Housing Program from the other 2. Head disability was the major contributing factor to distinguishing between the Housing Choice Voucher Program and Multifamily Housing. Families whose head’s that did not have a disability tended to be more likely to be in Multifamily Housing. Being an Asian, Black, Mixed race, or Hawaiian/Pacific Islander and non-Hispanic were also major contributing factors to being in the Multifamily Housing program rather than the Housing Choice Voucher Program.

**Linear Regression:**



In this linear regression model, I compared the “gross\_rent\_amnt” with the “asstn\_pymnt\_amnt”, dividing up the data by the household head’s race and year. This reveals how there is a certain limit to the amount of assistance a household gets in comparison to their overall rent, and the amount of deviation on the line of best fit from this limit highlights how a significant number of households do not receive the maximum amount of assistance. In addition, there is a great disparity between the number of households of the various races, as there are significantly more households whose heads are white or black compared to Asian or Hawaiian/Pacific Islander. In addition, another regression model was generated comparing the “asstn\_pymnt\_amnt” with the year for each of the races. Using the model, prediction could be made regarding the average amount of assistance a household given the household head’s race and a year. Using this model, we calculated that in 2023, the average white household will receive $694, the average Asian household $1168, the average black household $742, the average Native American household $725, the average Hawaiian/Pacific Islander household $1003, and the average mixed race household $962. It is unclear why the average Asian household is projected to receive the most amount of assistance, but this could be due to the group having fewer data points compared to white or black households.

**Cluster Analysis:** The Confusion Matrix (CM) can be interpreted As the columns being the number of instances within each of the grouped clusters determined by the k-means algorithm and the rows being how many instances of each ‘quality’ type are within each of the clusters. K-means clustering algorithm found 3 clusters of sizes 1: 17928 - Public Housing, 2: 4208 - Housing Choice Voucher Program, 3: 37019 - Multi-Family Housing. The implication is that the unsupervised machine learning model is predicting that more recipients are more closely related to being in Cluster 3 - Multi-Family Housing program rather than the Housing Choice Voucher Program as discovered in the exploratory data analysis. This may suggest that in the future HUD may find more recipients may qualify for Multi-Family Housing rather than the Housing Choice Voucher Program.



# RECOMMENDATIONS AND CONCLUSIONS

Based on the decision tree, it seems that people with disabilities are less likely to be in the Multifamily Housing program. We would recommend examining why families with heads with disabilities do not use the Multifamily Housing program as frequently and perhaps adjusting the program to be more accommodating for those with disabilities if deemed necessary.

It was also noted that Asians, Blacks, Mixed Race and Hawaiian/Pacific Islanders were more likely to be in Multifamily Housing. We would be curious to know whether they are overrepresented in this category or underrepresented in the Housing Choice Voucher Program and the reasons for this. If for example, these groups are being underrepresented in the Housing Choice Voucher Program, what are the reasons for this, and how can HUD make this particular program more available to these particular racial groups (and vice versa, if Whites and Native Americans are underrepresented in the Multifamily Housing program, what are the reasons and what steps can be taken to make this program more accessible to members of these populations who need it).

On average the cluster averages were close to the averages found when conducting exploratory data analysis, because Housing Choice Voucher Program (Cluster 2) was much smaller than what was found in exploratory data analysis, the correlation cannot be antiquated. The model suggested more families in the future to receive multi-family housing benefits which may suggest that the people more people in a household than averages found in exploratory data analysis may qualify for multi-family housing rather than housing choice vouchers.

The linear regression model determined rent prices to increase in the next five years; however, found a positive relationship between increasing rent costs and assisted payments costs suggesting and validating that assistance payments by HUD to cover a similar ratio as found today. However, the linear model predicted that various races and ethnicities will receive marginally different assistance amounts. The margin of errors found with all ethnic groups outside of the linear regression line implies differences among races and ethnic groups.

To support the pillars of sustainability the team recommends:

1. Comparing existing business rules for each assistance program to the results found in the decision tree to determine if commonalities and correlations exist between the two
2. Conducting a Decision Tree Analysis analogous to each program type and comparing to existing business rules for each program
3. Perform an independent analysis that identifies and statistically validates why differences exist among ethnic groups and races in terms of assistance amounts by HUD to eliminate any suggestions of demographic bias.
4. Link geographic data to the dataset to uncover an hidden insights or discoveries between income, rent, and assistance payments by HUD by geographic region/location.

# ACKNOWLEDGMENTS

Thanks to Data Challenge 2020: all the volunteers and all Data Challenge 2020 staff and mentors (Dr. Elizabeth Bensignore) who provided support for the workshops. Thanks also to the HUD advisors (Jacquie) and sponsors who answered the team’s questions.

# REFERENCES

1. HUD Strategic Plan FY2018-2022, HUD Strategic Plan FY2018-2022 (2019). Retrieved from <https://www.hud.gov/sites/dfiles/SPM/documents/HUDSTRATEGICPLAN2018-2022.pdf>
2. Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques. Waltham: Morgan Kaufmann.