**Summary of Result**

**Executive Summary**

* Data were cleaned and merged in Pandas, Python. Since Missing/NA data in all column were very small. They were removed. Only Veterans Column missing value was not removed.
* Column Age Group, Service Amount Sum, Service Number Total were created. Service Amount Sum, Service Number Total column received 0 value for Intake Survey.
* B1cIllegalDrugs was used for short term effect and E4CommitedCrime were used for midterm effect and D3EmploymentSituationId were used for long term effect.
* Mean, median was calculated for target variables, D2EducationLevelId, ServiceAmt sum ServiceNum Total, age. For Ethnorace, Veteran, Gender unique value counts were calculated. Tables are provided in second page.
* We applied t-test to detect the difference between mean of target variables between Intake and Post survey. P-value of t-test was significant for 3 variables.
* The regression model was performed for each of 3 target variables. Cross Validation were performed for the model validation.
* **The Predictors for regression model were:** InterviewIntake, Gender, Ethnicity, Ethnorace, Veteran, AgeGroup, ServiceAmt sum, ServiceNum Total
* Out model have shown that program was effective in predicting target variable.
* For **B1cIllegalDrugs:** treatment, Gender Male, Veteran unknown, Age Group Old, Age Group Young, Service Amt sum were significant predictors.
* For **E4CommitedCrime:** treatment, Gender Male, Age Group Old, Age Group Young were significant predictors.
* For **D3EmploymentSituationId:** treatment, Gender Male, Ethnicity Not Hispanic or Latino, Ethnorace Black or African American, Ethnorace Hispanic , Ethnorace Multi Cultural, Ethnorace Native American or Alaskan Native , Ethnorace unknown, Ethnorace White , Veteran unknown, Veteran Yes , Age Group Old were significant predictors.
* Further analysis is required for determining the effect of the program on each group and specifically what type of services is most effective.
* Our regression model had a low adjusted R squared. There is large amount of variance not explained by this model. More predictive analysis is required to 1) perform better feature engineering, 2) find a better model to fit, 3) perform model validation and analysis.
* There is large amount of missing data in Veteran Column which can have significant effect in the model.

**Data cleaning**

Data was imported to python. We used Pandas to clean and merge the datasets.

Missing value was checked in Survey, Client and Service datasets.

Since number of missing/NA columns are very small. Missing values and [-1,-7, -8, -9, 'Missing'] values were removed in target variables in Survey dataset. But, in a rigorous study we might need to impute them since we might be having response bias.

missing values on veterans Colum have some different characteristics at least on one target variable of interest. We will keep it as an 'unknown' category as it might help with our model later on and there is no explicit reason to drop these missing values as they are around 60% of total number of rows. we dropped the rest of the columns missing value since their size is too small.

We merged the three data set and created one large client demographic dataset.

ServiceAmt sum and ServiceNum Total column was created based on Service Data. Service Amount Sum, Service Number Total column received 0 value for Intake Survey for further analysis.

Age group Column was also created based on DOB column.

We chose 3 target variables to look at the short term, mid-term and Long-term effect of program.

B1cIllegalDrugs was used for short term effect and E4CommitedCrime were used for midterm effect.

and D3EmploymentSituationId were used for long term effect.

**Data analysis and exploration**

**D2EducationLevelId ServiceAmt sum ServiceNum Total age**

**min** 0.000000 0.000000 1.000000 20.000000

**max** 18.000000 3475.000000 74.000000 100.000000

**median** 12.000000 806.500000 16.000000 40.000000

**mean** 12.211842 862.935313 17.350011 42.431633

**sem** 0.020441 6.122404 0.119356 0.109915

**E4CommitedCrime D3EmploymentSituationId B1cIllegalDrugs**

**min** 0.000000 0.000000 0.000000

**max** 400.000000 7.000000 30.000000

**median** 0.000000 3.000000 0.000000

**mean** 2.122286 2.733223 2.248430

**sem** 0.049014 0.011820 0.045057

**Ethnorace Veteran Gender**

**unknwon** 3471 **unknwon** 7599 **Female** 7225

**White** 3124 **No** 3407 **Male** 3971

**Hispanic** 2540 **Yes** 192

**Native American or Alaskan Native** 1231

**Black or African American** 486

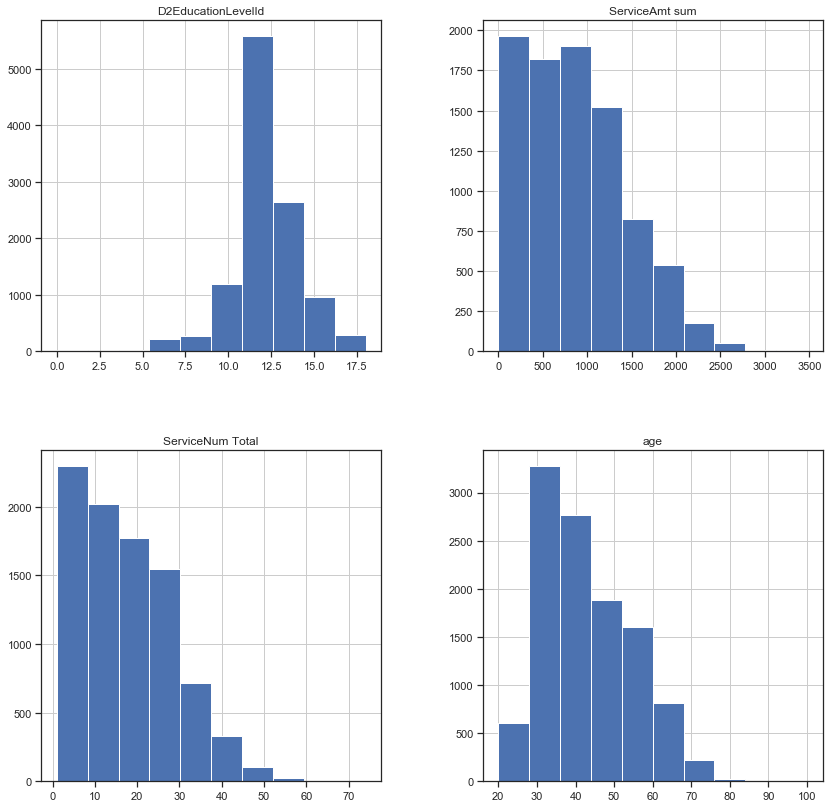
**Multi Cultural** 290

**Asian** 31

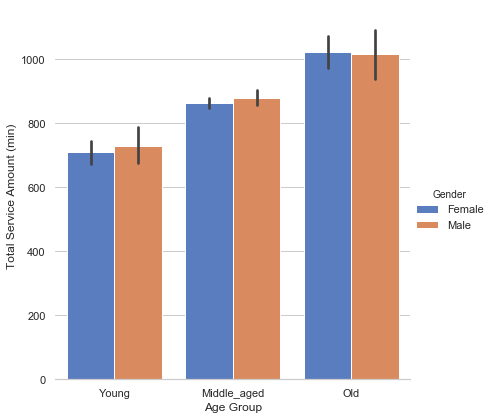
**Hawaiian or Pacific** 25

**Data visualization**

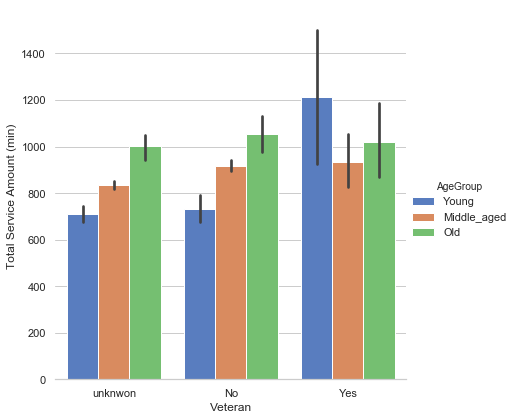
**Distribution of data in variables age, ServiceAmt sum, ServiceNum Total, EducationLevel.**

****

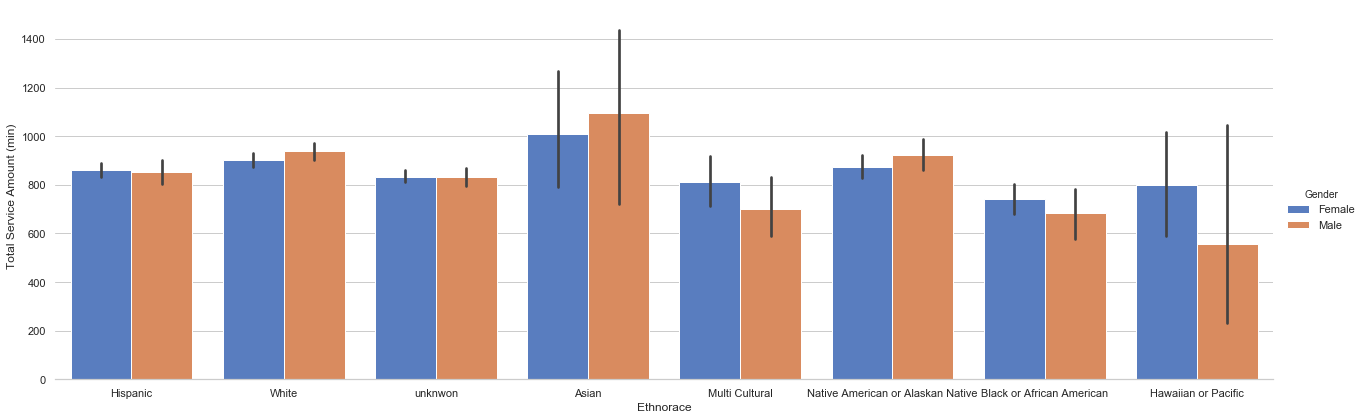
**Histogram of agegroup by Total Service Amount**

****

**Histogram of Veteran by Total Service Amount**

****

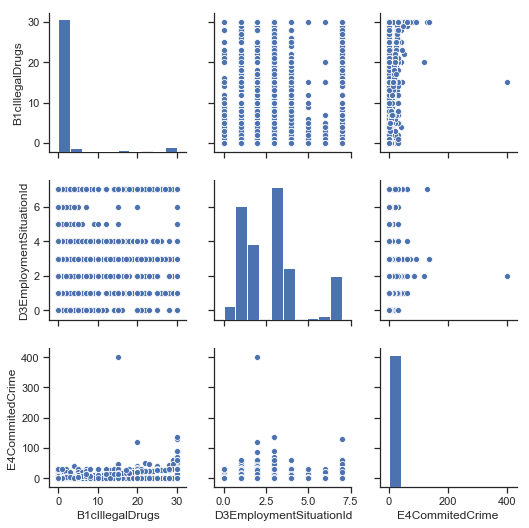
**Histogram of Ethnorace by Total Service Amount**

****

**Data modeling**

We used R for the regression modeling.

The correlation of these variables were checked before the predictive modeling.



We applied t-test to detect the difference between mean of target variables between Intake and Post survey. P-value of t-test was significant for 3 variables.

The regression model was performed for each of 3 target variables, B1cIllegalDrugs, D3EmploymentSituationId and E4CommitedCrime. Cross Validation were performed for the model validation.

The Predictors for regression model were: InterviewIntake, Gender, Ethnicity, Ethnorace, Veteran, AgeGroup, ServiceAmt sum, ServiceNum Total.

Our model has shown that program was effective and could predict the target variable.

**For B1cIllegalDrugs:** treatment, Gender Male, Veteran unknown, Age Group Old, Age Group Young

, Service Amt sum were significant predictors.

**For E4CommitedCrime****:** treatment, Gender Male, Age Group Old, Age Group Young were significant predictors.

**For D3EmploymentSituationId:** treatment, Gender Male , Ethnicity Not Hispanic or Latino, Ethnorace Black or African American, Ethnorace Hispanic , Ethnorace Multi Cultural, Ethnorace Native American or Alaskan Native , Ethnorace unknown, Ethnorace White , Veteran unknown, Veteran Yes , Age Group Old were significant predictors.

**Future Direction**

Further analysis is required for determining the effect of the program on each group and specifically what type of services is most effective.

Our regression model had a low adjusted R squared. There is large amount of variance not explained by this model. More predictive analysis is required to 1) perform better feature engineering, 2) find a better model to fit, 3) perform model validation and analysis.

There is large amount of missing data in Veteran Column which can have significant effect in the model.

* **mtry**: Number of variables randomly sampled as candidates at each split.
* **ntree**: Number of trees to grow.

root mean square error (RMSE) to estimate how well our random forest was able to predict our test set outcomes