**ADAlytix – EDA Phase I**

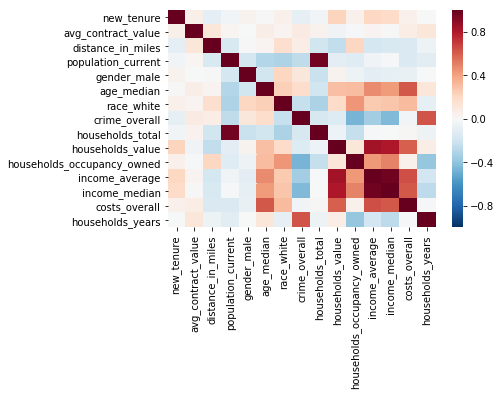
The results summarized below make use of the clean aggregated data in “data\_final.csv”. This data consists of information averaged out for 549 zip codes, across 15 variables.

**Derived variable(s):**

The variable “households\_yearbuilt” gives the year when households were first built in that zip code. This variable has been used to create a new variable “households\_years” which gives the total years that have elapsed since households were built in a zip code.

**Correlation analysis:**

The following plot shows the correlation among all the variables found:



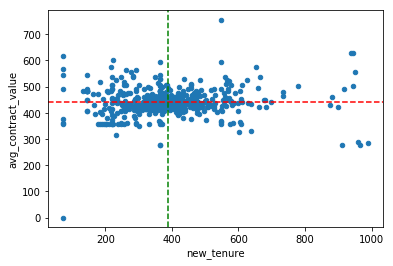
Though the variables of interest “new\_tenure” and “avg\_contract\_value” show no direct correlation with any variables, several important dependencies could be observed:

* “crime\_overall” shows a 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.

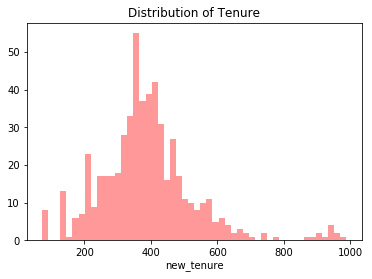
The correlation observed here can be used to remove redundancy in variables and potential confounding variables like “crime\_overall”.

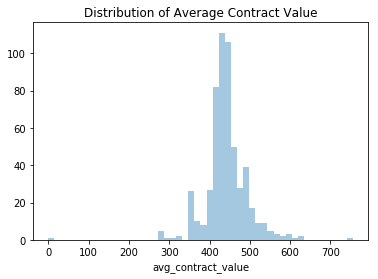
**Contract value vs Tenure:**

The following plot shows the trend between “avg\_contract\_value” and “new\_tenure”, where the vertical and horizontal lines indicate the mean.



The above plot is indicative of outliers or records with extremely high and low “avg\_contract\_value” and “new\_tenure”. Examining the behaviour of these outliers would be vital for gathering insights. The distribution plots of “avg\_contract\_value” and “new\_tenure” are shown below:





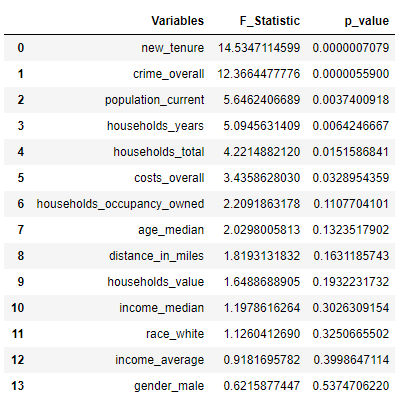
It is seen that both “new\_tenure” and “avg\_contract\_value” follow normal distributions. Hence, records on the lower and upper bounds can be safely captured using the bounds mean+(2 times SD) and mean-(2 times SD).

* Mean of value = 441, SD of value = 53
* Mean of tenure = 389, SD of tenure = 143

**ANOVA:**

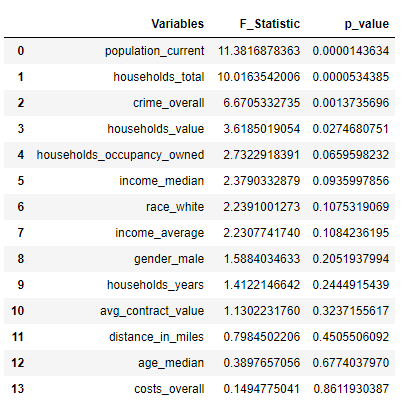
To perform ANOVA , the zips have been grouped into three – low value zips (10), mid value zips (525) and high value zips (14), based on the “avg\_contract\_value”. Based on “new\_tenure”, the following groups are observed – low tenure zips (8), mid tenure zips (522) and high tenure zips (19).

The ANOVA results for “avg\_contract\_value” have been summarised below. The null hypothesis is that “avg\_contract\_value” has no impact in each of the other variables. Hence, the null hypothesis expects that means of each variable for the entire populations of low value zips, mid value zips and high value zips must not be different. The null hypothesis is rejected for the following variables: “new\_tenure”, “crime\_overall”, “population\_current”, “households\_years”, “households\_total” and “costs\_overall”.



“crime\_overall” has a high significance value. However, “crime\_overall” might be found to have high significance because of the presence of other variables with which it is correlated – like “population\_current” and “households\_total”. The presence of “households\_years” can be considered derivative to the effect of “crime\_overall” as both are highly correlated.

Similarly, the ANOVA results of “new\_tenure” has been tabulated below. “population\_current”, “households\_total”, “crime\_overall” and “households\_value” are the variables with high significance levels. “crime\_overall” may appear here due to a similar reason as discussed above.



The overlay plots of distributions for each of the variables can be found in the attached Jupyter notebook.