Data Mining Proj

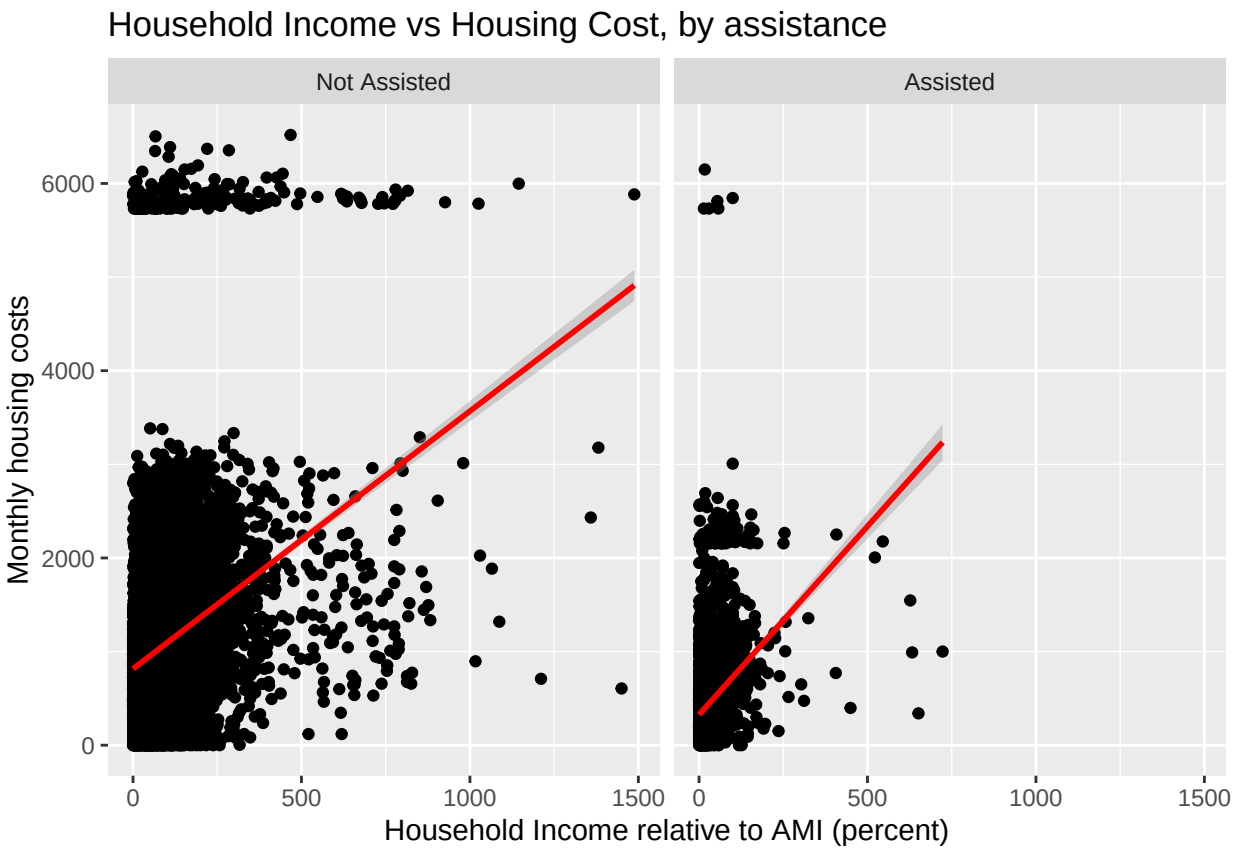
#Exploratory Data Analysis

##Exploration with Plots

This first plot exemplifies our problem definition, to attempt to identify characteristics of households that are likely in need of assistance, but are not recieving it. The plot compares income (relative to Area Median Income) to housing costs (at median interest relative to area median income). The plots are broken into households that receive assistance vs those that do not. But unfortunately the plots are not particularly different and have quite a bit of overlap.

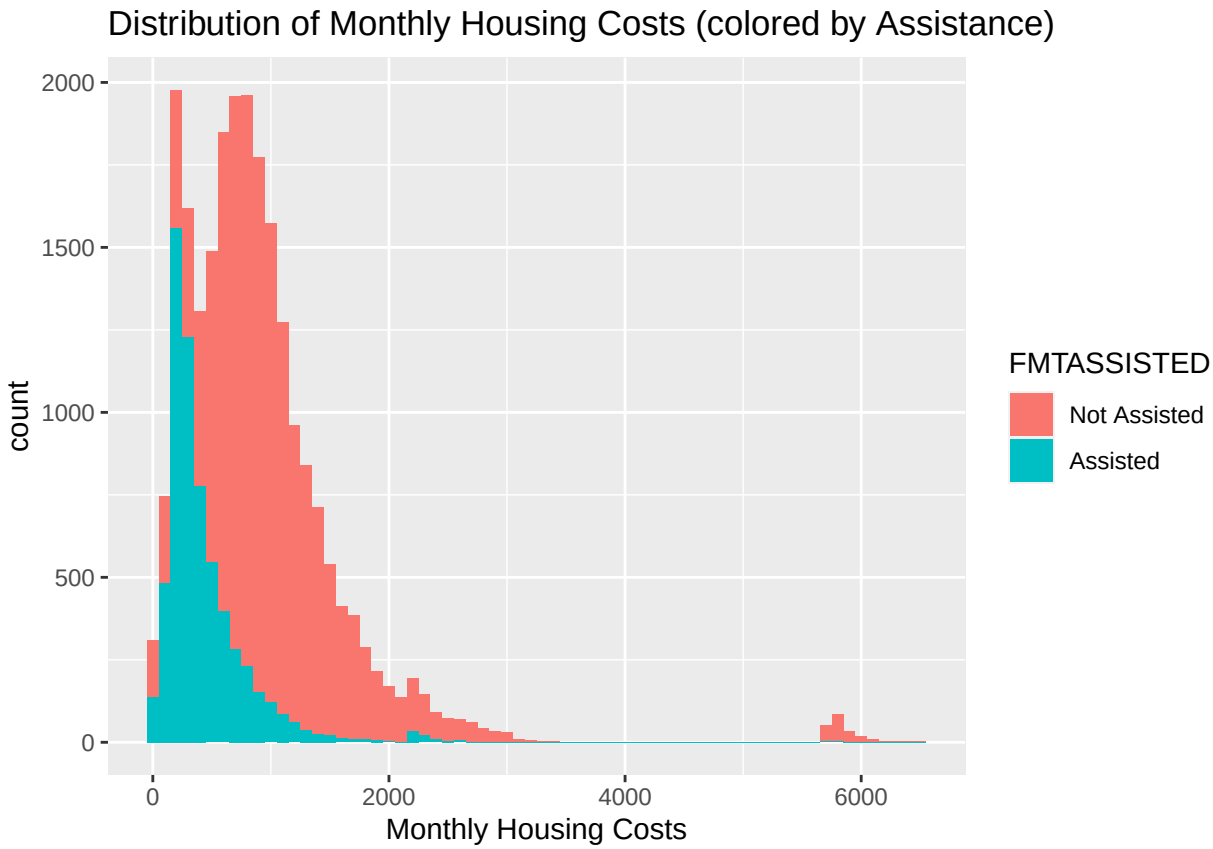
```
ggplot(data = housingClean, mapping=aes(x=INCRELAMIPCT , y=ZSMHC )) + geom_point() + geom_smooth(method
  facet_wrap('FMTASSISTED') +
  labs(x ='Household Income relative to AMI (percent)' , y = 'Monthly housing costs', title = 'Household
```

```
## `geom_smooth()` using formula 'y ~ x'
```



In that plot we can see that there are a few households that are behaving as outliers. To highlight this, below we have plotted a histogram of housing costs (stacked with the assisted variable for the color).

```
#histogram of housing costs
ggplot(data = housingClean, mapping=aes(x= ZSMHC, fill = FMTASSISTED)) + geom_histogram(binwidth = 100)
```



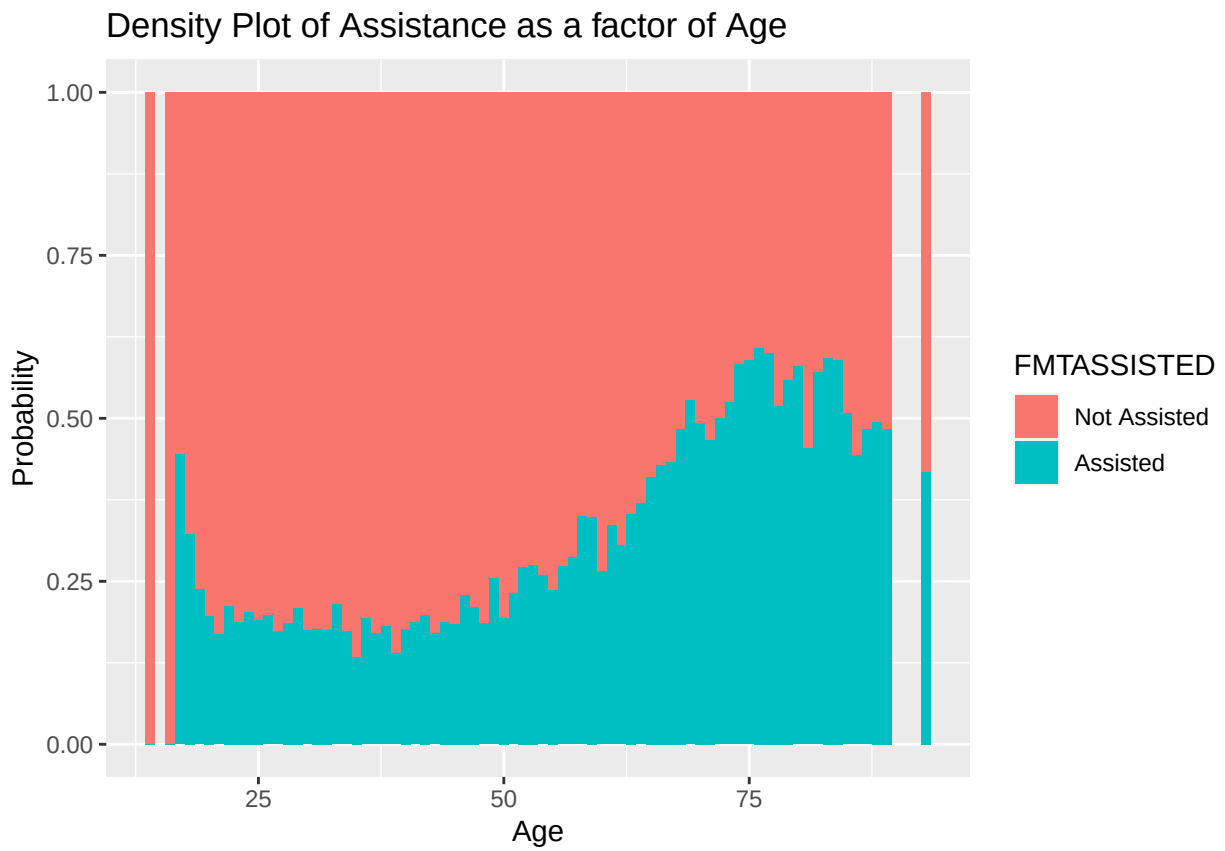
We decided to remove these outliers for the rest of our analyses.

```
# restrict data to households with monthly housing costs under $4000
housingClean = housingClean[which(housingClean$ZSMHC < 4000),]
```

```
#Age vs Assistance
ggplot(housingClean, aes(AGE1, fill = FMTASSISTED)) + geom_histogram(binwidth = 1, position = "fill") +
```

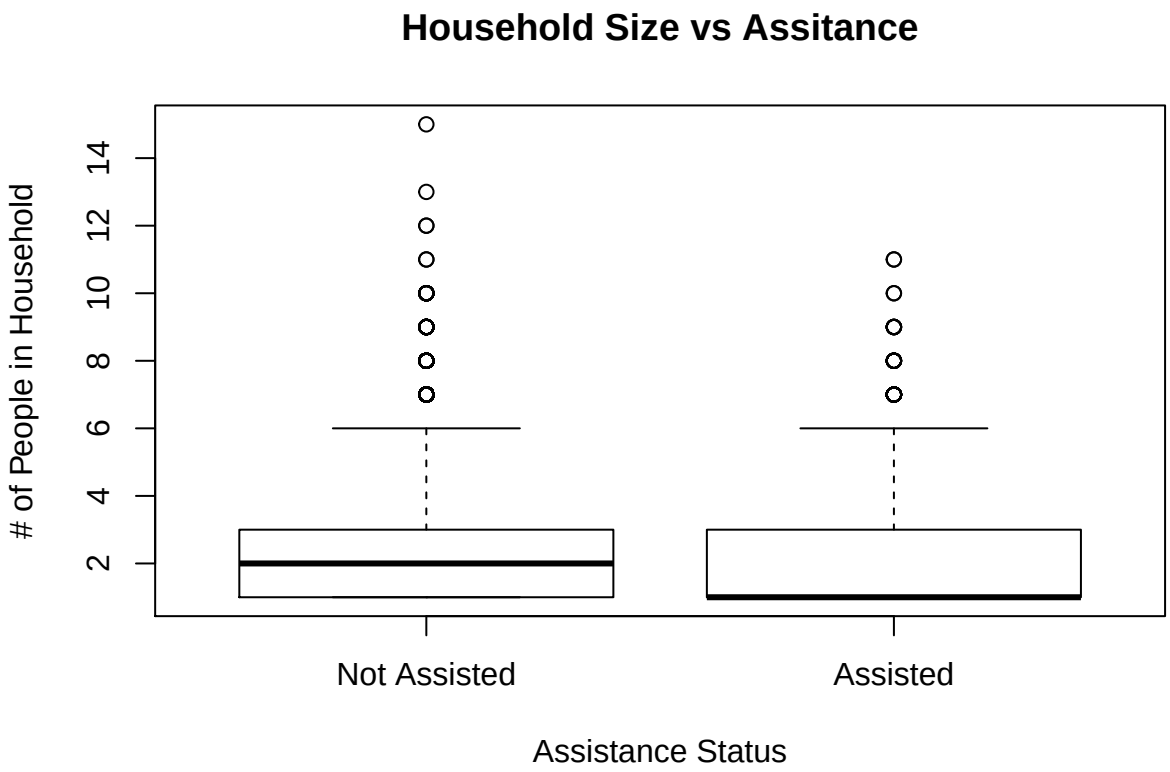
How Age affects whether a household is assisted:

```
## Warning: Removed 8 rows containing missing values (geom_bar).
```



This plot shows that people are more likely to receive housing assistance when there are very young adults (about 16-17 years old) and after the age of about 65 year old.

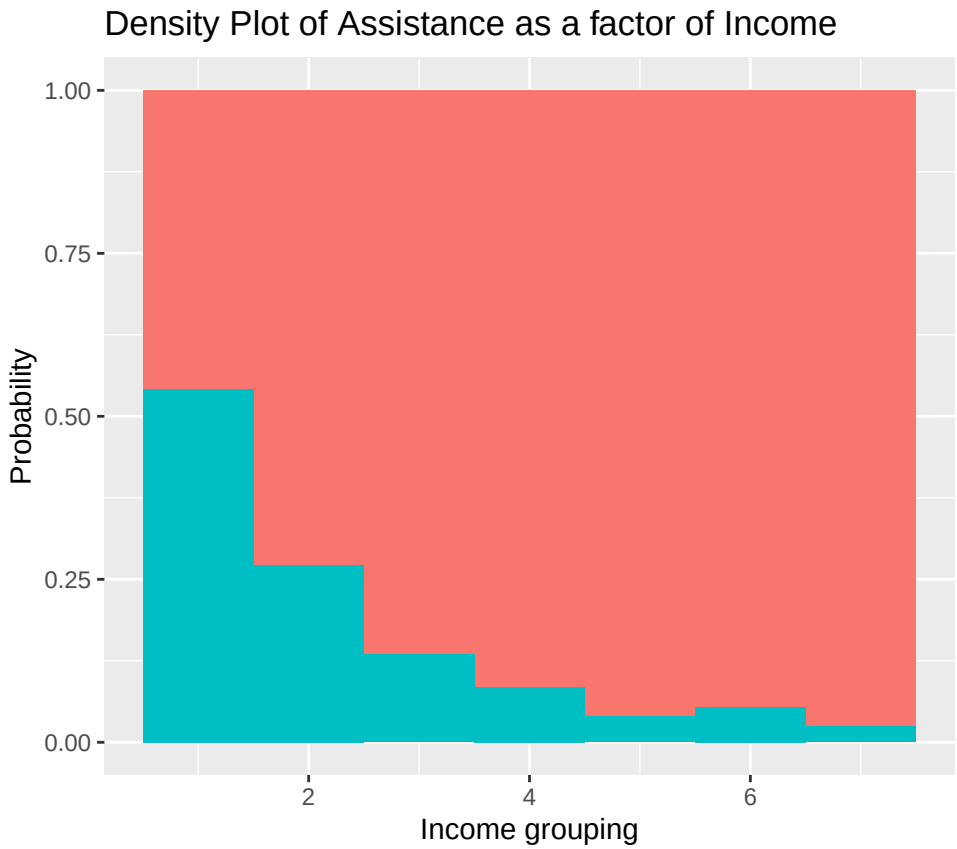
```
# Box plot of Number of People in household vs Assisted
boxplot(PER~FMTASSISTED,data= housingClean, main="Household Size vs Assitance",
   xlab="Assistance Status", ylab="# of People in Household")
```



Size of Household

This plot shows the distribution of the number of people in the households, broken up by assistance status. There are interesting insights that contradict our assumptions about assisted households, in that they tend to be smaller (just 1 person) than unassisted households (2 people). This may be a result of two or more people living together having an easier time pooling resources to share housing costs. Other insights are unclear without deeper analysis.

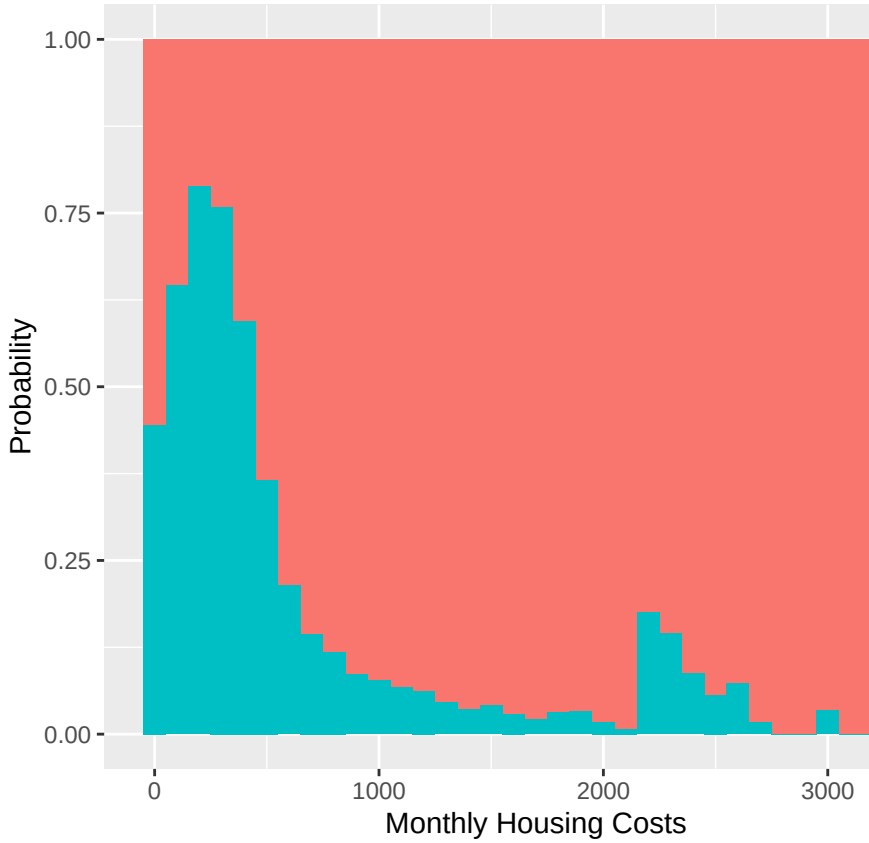
```
# Income (HH Income relative to AMI, categorized) vs Assisted
#make subset of uncleaned data to use income categorization (and Housing cost categorization in the fol
df= subset(housingData, FMTASSISTED != ".")

ggplot(df, aes(INCRELAMICAT, fill = FMTASSISTED)) + geom_histogram(binwidth = 1, position = "fill") + la
```



Income affects of Assistance status

This plot shows that as households incomes increase, the likelihood that they receive assistance is lower. Which is quite intuitive.

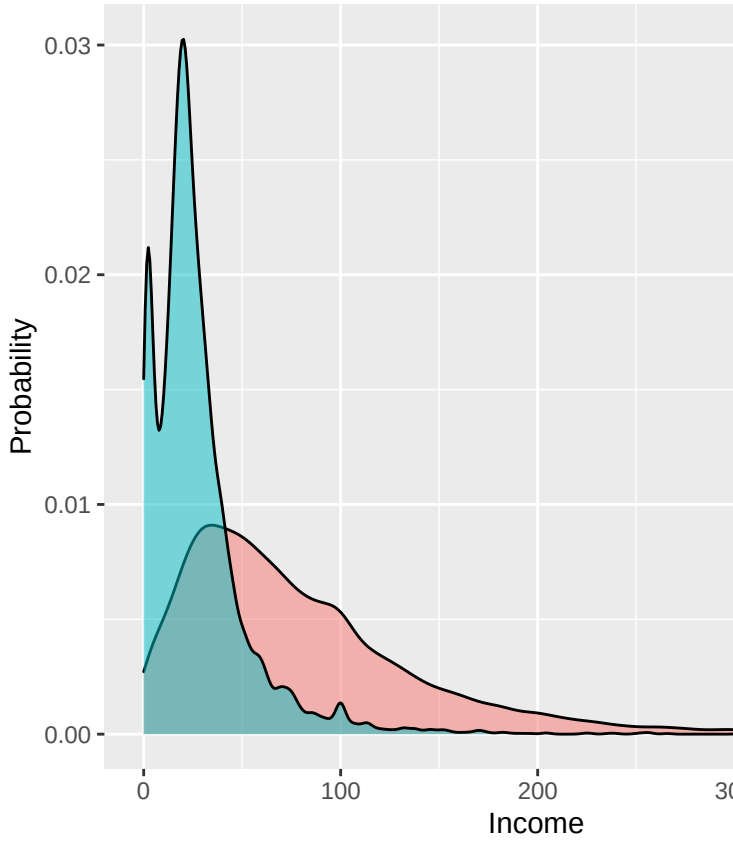
```
ggplot(housingClean, aes(ZSMHC, fill = FMTASSISTED)) + geom_histogram(binwidth = 100, position = "fill"
```



Housing Cost affects on Assistance status

This plot shows the trend that households with lower housing costs are more likely to receive assistance.

```
#restrict data to INCRELAMIPCT < 400 to make plot easier to view
df = housingClean[which(housingClean$INCRELAMIPCT < 400),]
ggplot(df, aes(INCRELAMIPCT, fill = FMTASSISTED)) + geom_density(alpha=0.5) + labs(x = 'Income', y='Prob
```



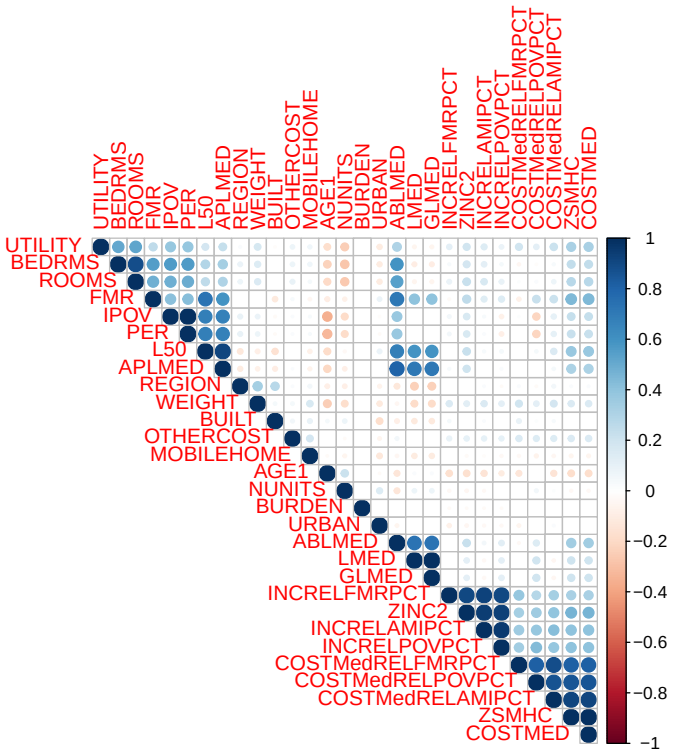
Income distribution broken up by Assistance Status
This plot shows that households with lower income are more likely to be assisted.

Besides the plots shown, other exploratory analyses included the following, but did not show any particular insights or useful interactions:
Age vs Monthly Housing Cost
MOBILEHOME status vs Assistance
Housing Adequacy vs Assistance
Housing Adequacy vs Income
Urban status vs Assistance
Urban status vs Income and Housing Costs

Assessment of Mulit-collinearity

To eliminate redundant variables, and prepare the data for logistic regression, we checked all continuous variables in our "cleaned" data set for collinearity.

```
#restrict data to only continuous variables
cont.data = housingClean[,c(1:27,34,35)]

#create correlation matrix (with p-value matrix)
corrmat <- rcorr(as.matrix(cont.data))

# Create corrlation plot, insignificant correlations are left blank
corrplot(corrmat$r, type="upper", order="hclust",
         p.mat = corrmat$P, sig.level = 0.01, insig = "blank")
```



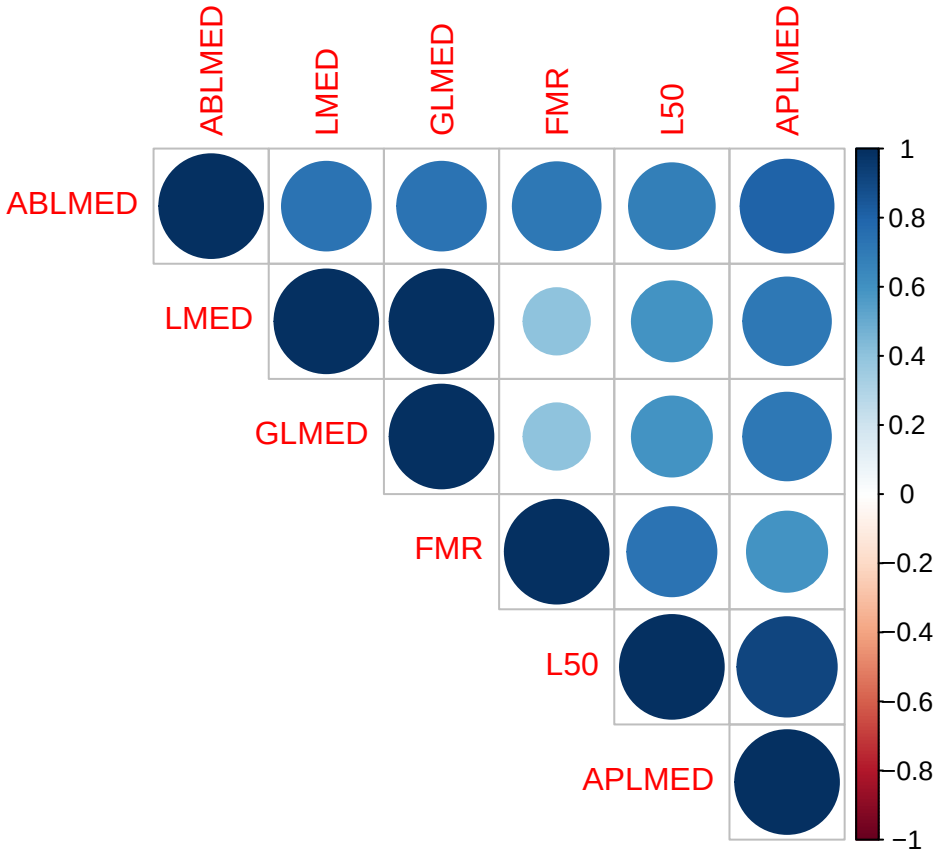
After computing correlation coefficients of all variables, we discovered two pairs of collinear variables:
IPOV + PER = 1.00
BEDRMS + ROOMS = 0.99

as well as three groupings of collinear variables, which are each visualized, with specific coefficients listed.

Variables related to Area median income and Fair Market Rate

FMR, L50, LMED, GLMED, APLMED, ABLMED

```
#restrict data to "Median Income" related variables
lmed.data = housingClean[,c('FMR', 'L50', 'LMED', 'GLMED','APLMED','ABLMED')]
lmed.corr <- rcorr(as.matrix(lmed.data))

# Plot correlation viz where insignificant correlations are left blank
corrplot(lmed.corr$r, type="upper", order="hclust", p.mat = lmed.corr$P, sig.level = 0.01, insig = "bla
```



LMED + GLMED = 1.00
FMR + L50 = 0.89
FMR + APLMED = 0.86
FMR + ABLMED = 0.91
L50 + APLMED = 0.98
L50 + ABLMED = 0.89
APLMED + ABLMED = 0.93
APLMED + GLMED = 0.71
APLMED + LMED = 0.71
LMED + ABLMED = 0.73
LMED + FMR = 0.41
LMED + L50 = 0.59

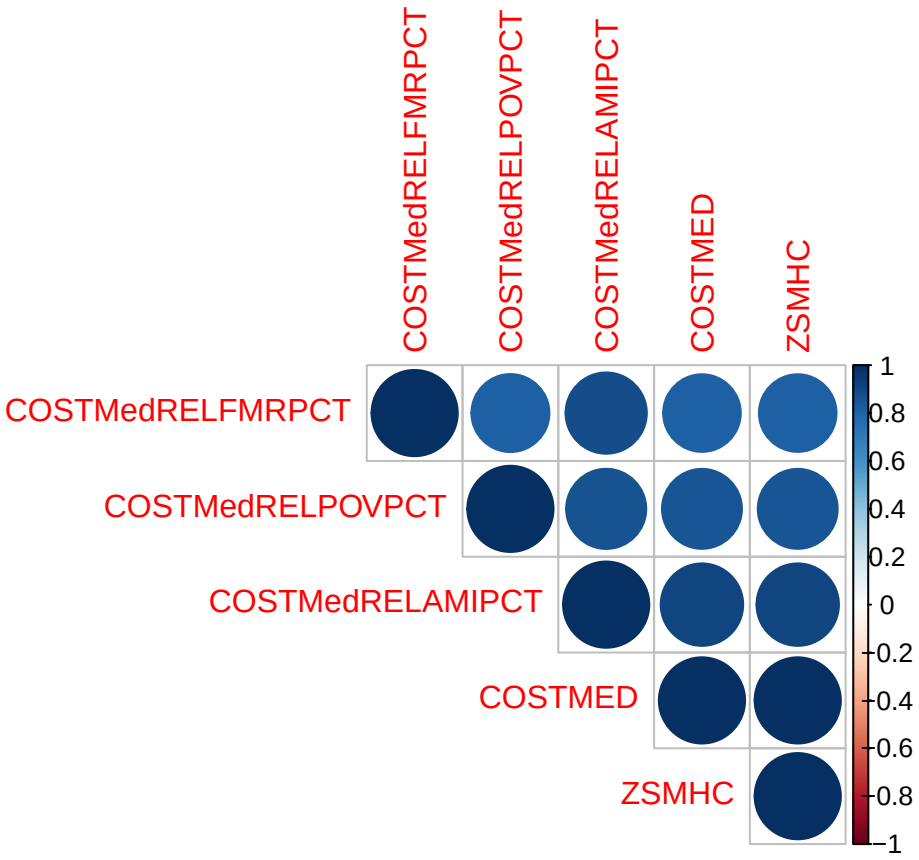
From this grouping we chose to keep `APLMED` (Median Income Adjusted for # of Persons) because it is representative of this whole grouping of collinear variables, since it is the most correlated with other variables in the group. We also decided to keep `FMR` because it is reasonably different in its definition, and its coefficient with `APLMED` of 0.86 is not quite over the threshold of 0.9. We also chose to keep `L50` for similar reasons, as its definition is also different.

Housing Cost Variables

`COSTMED`, `COSTMedRELAMIPCT`, `COSTMedRELPOVPCT`, `COSTMedRELFMRPCT`, `ZSMHC`

```
#restrict data to "Cost" related variables
costmed.data = housingClean[,c('COSTMED', 'COSTMedRELAMIPCT', 'COSTMedRELPOVPCT', 'COSTMedRELFMRPCT', 'Z
costmed.corr <- rcorr(as.matrix(costmed.data))

# Plot correlation viz where insignificant correlations are left blank
```

```
corrplot(costmed.corr$r, type="upper", order="hclust", p.mat = costmed.corr$P, sig.level = 0.01, insig =
```



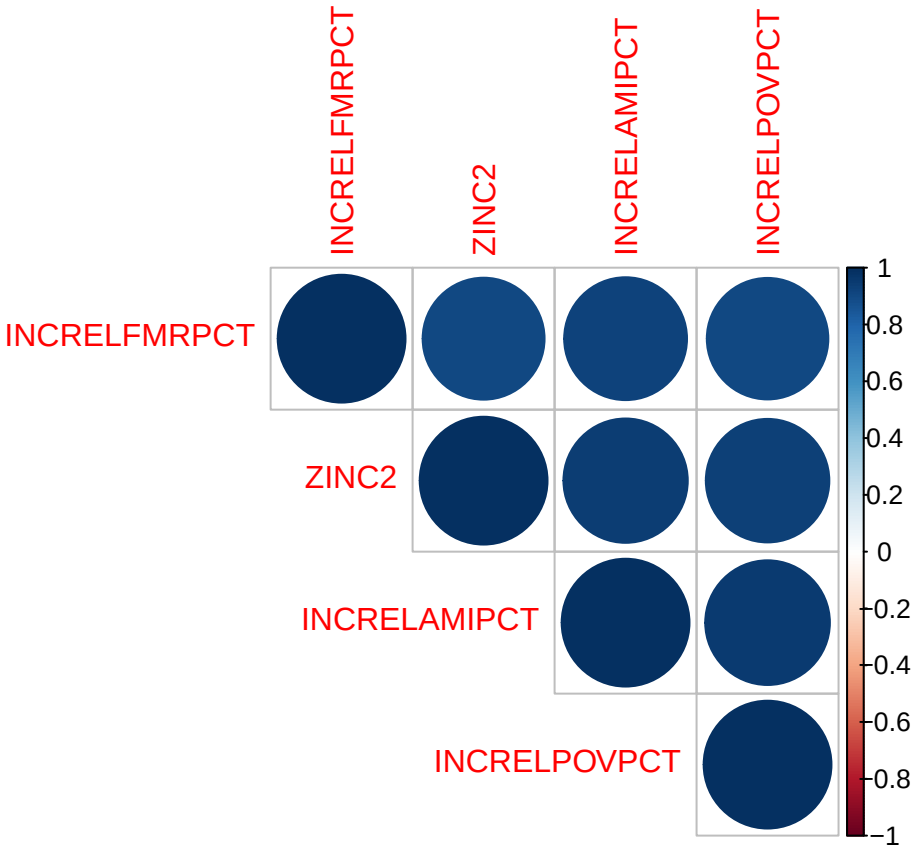
ZSMHC + COSTMED = 1.00
ZSMHC + COSTMedRELAMIPCT = 0.96
ZSMHC + COSTMedRELPOVPCT = 0.93
ZSMHC + COSTMedRELFMRPCT = 0.92
COSTMED + COSTMedRELAMIPCT = 0.96
COSTMED + COSTMedRELPOVPCT = 0.93
COSTMED + COSTMedRELFMRPCT = 0.92
COSTMedRELAMIPCT + COSTMedRELPOVPCT = 0.96
COSTMedRELAMIPCT + COSTMedRELFMRPCT = 0.98
COSTMedRELPOVPCT + COSTMedRELFMRPCT = 0.96

From this grouping, we chose to keep `ZSMHC` because, as seen in the first exploratory plots, this variable shows a stark gap between most of the households, and a few outliers over $4000. Because of this we also chose to subset our data without these outlier points.

Income Variables

`ZINC2`, `INCRELAMIPCT`, `INCRELPOVPCT`, `INCRELFMRPCT`

```
#restrict data to "Income" related variables
income.data = housingClean[,c('ZINC2', 'INCRELAMIPCT', 'INCRELPOVPCT', 'INCRELFMRPCT')]
income.corr <- rcorr(as.matrix(income.data))

# Plot correlation viz where insignificant correlations are left blank
corrplot(income.corr$r, type="upper", order="hclust", p.mat = income.corr$P, sig.level = 0.01, insig =
```



ZINC2 + INCRELAMIPCT = 0.97
ZINC2 + INCRELPOVPCT = 0.97
ZINC2 + INCRELFMRPCT = 0.96
INCRELAMIPCT + INCRELPOVPCT = 0.98
INCRELAMIPCT + INCRELFMRPCT = 0.99
INCRELPOVPCT + INCRELFMRPCT = 0.98

From this grouping, we chose to keep `INCRELAMIPCT` because according to the data set documentation, housing cost relative to AMI is the most common standard used in affordability discussions of the three standards provided (fair market rent - FMR, area median income - AMI, and poverty-level income - POV).

Our final data subset

Based on our understanding of the data gained from exploratory analysis and collinearity checked, we have limited our data to 18 variables that we will use to fit predictive models.

```
housingClean = housingClean[,c("AGE1", "REGION", "FMR", "L50", "BEDRMS", "BUILT", "NUNITS", "PER", "ZSM
```