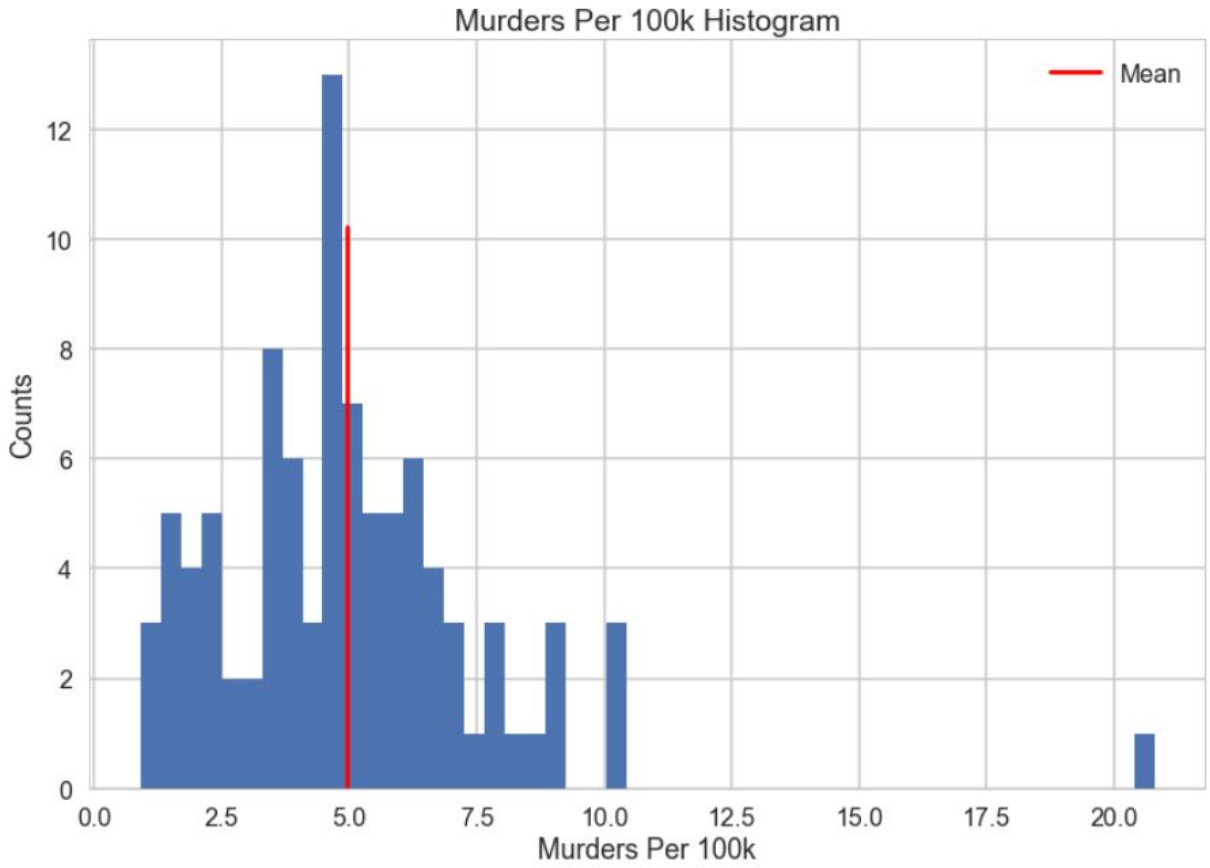
Milestone #3 - Group 14

1. **A description of the data: what type of data are you dealing with? What methods have you used to explore the data?**

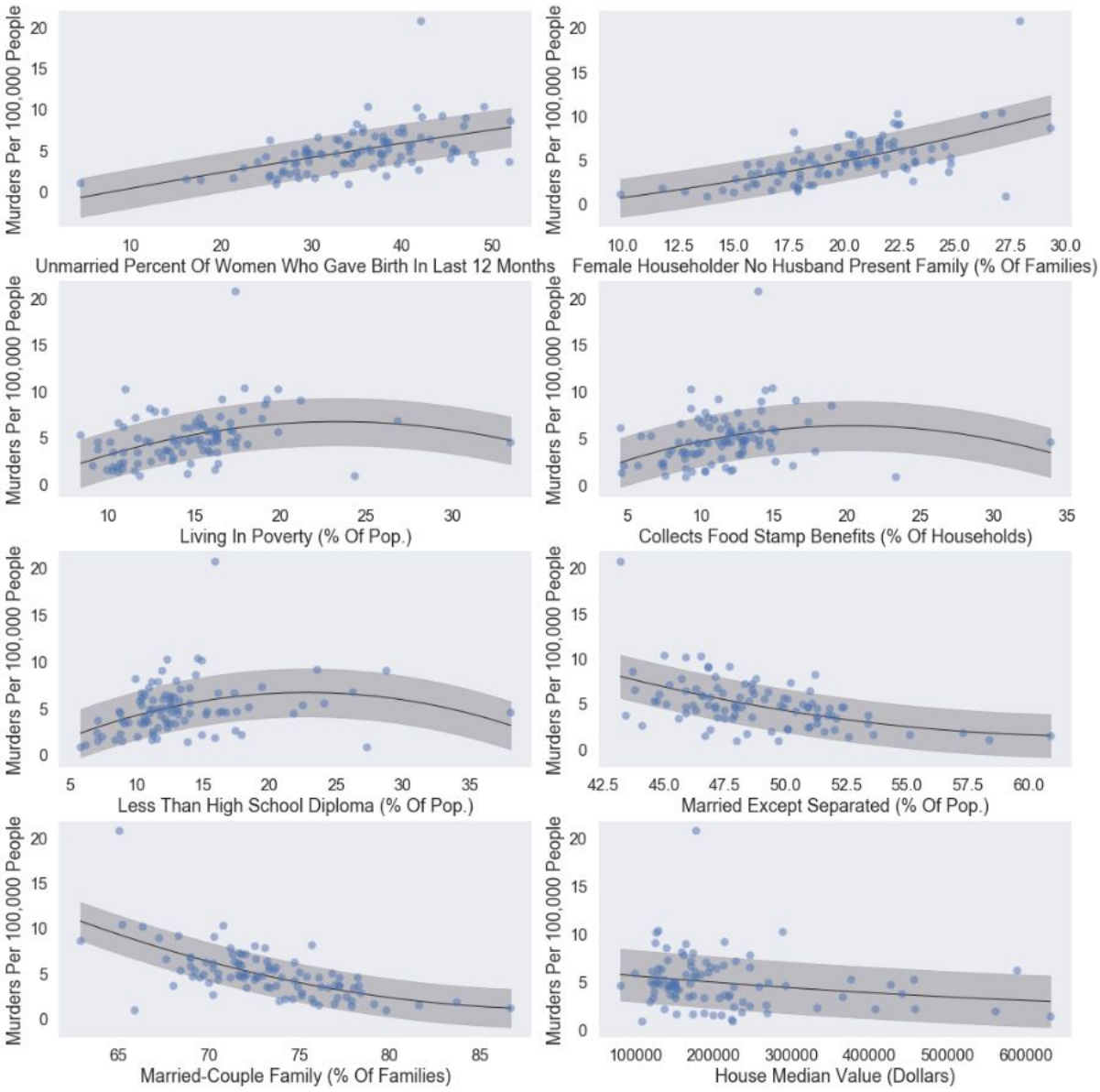
Data was retrieved from two sources. Murder rates were scraped from ucr.fbi.gov website by MSA, and census data was downloaded from factfinder.census.gov website and imported into Python. There are numerous inconsistencies in data between years so only 2010 was arbitrarily chosen to perform EDA first. Subset of features was selected and renamed for better readability. Once EDA is completed and final (5-10) features are chosen for further modeling, data for those features will be pulled for each year and combined into one dataset. Two datasets (fbi and census) are merged on MSA name, and only matching rows are considered for analysis.

For the EDA, we first plotted a histogram to check the distribution of the response variable. Then, we scatter-plotted ALL of the possible features against the murder rate in order to observe which were most relevant. From those, we found about 25 that were more relevant than the others, and we included the 8 most explanatory of those in an exhibit below.

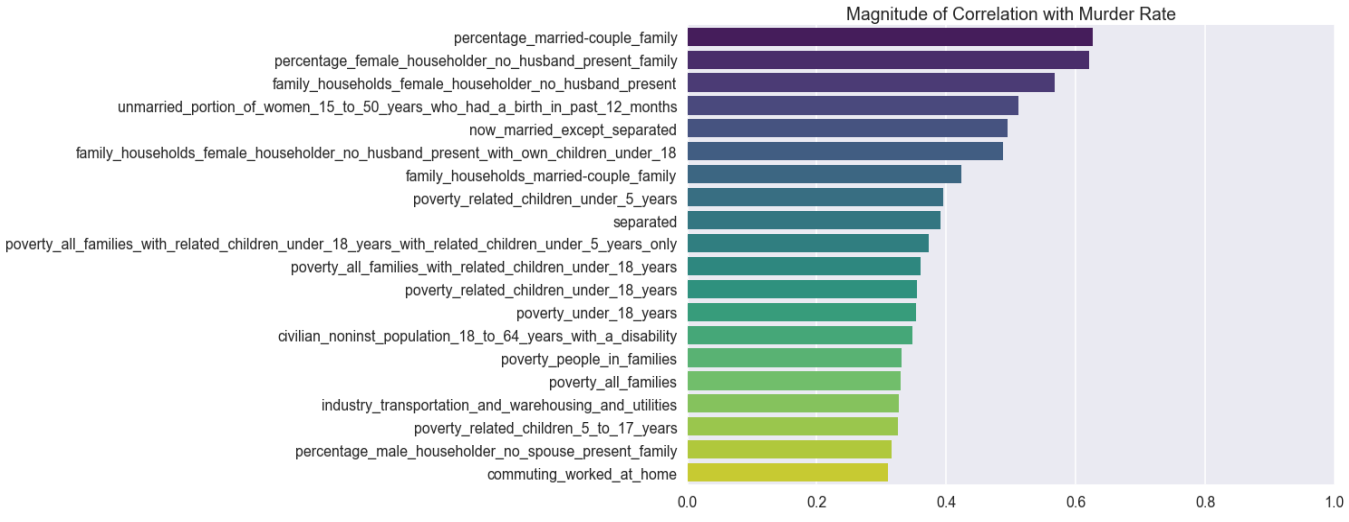


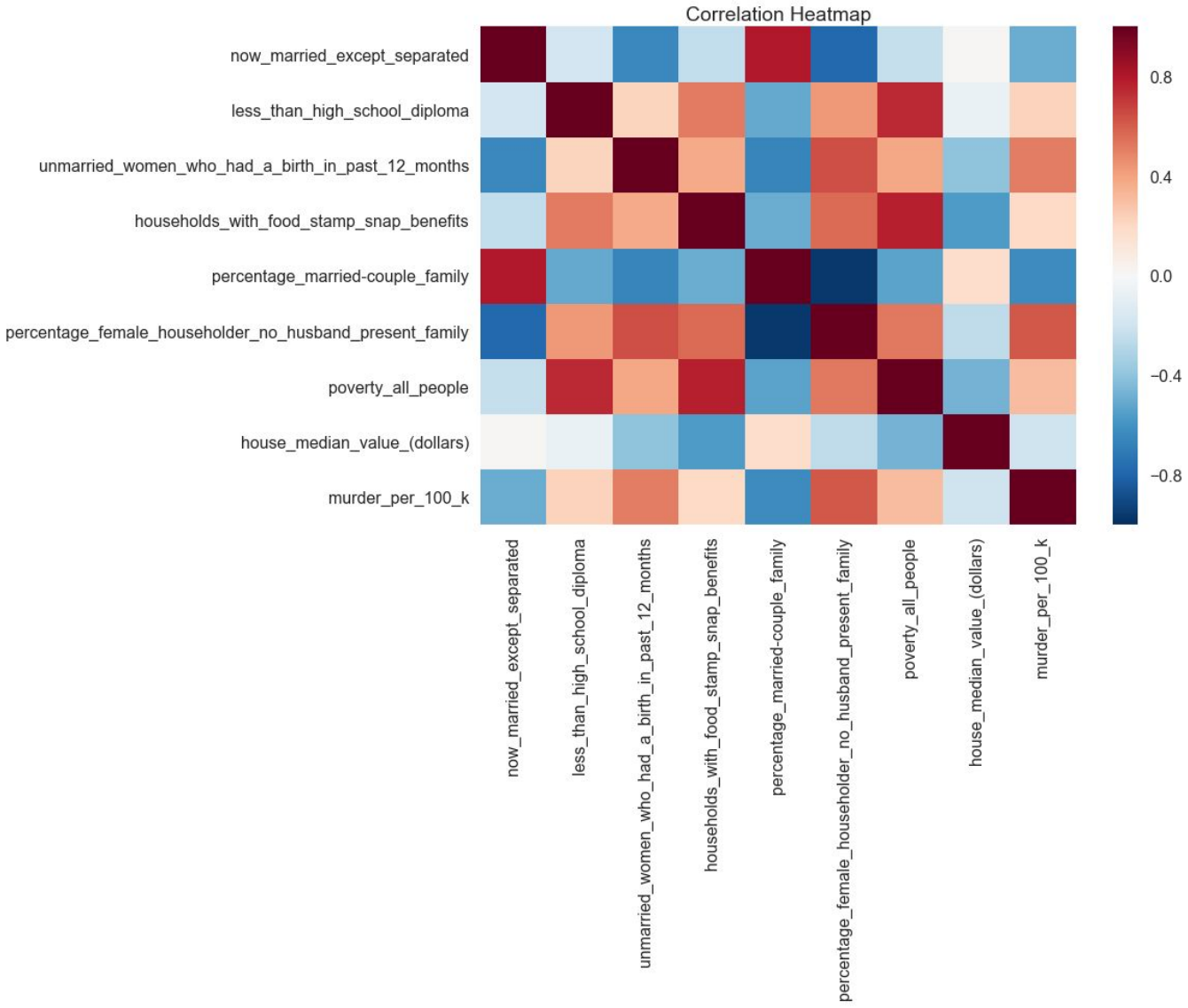
2. **Visualizations and captions that summarize the noteworthy findings of the EDA.**

The response variable looks somewhat normally distributed with a mean of 5 murders per 100,000 of population (per year). New Orleans was the lone very strong outlier, with a murder rate of over 20 per 100,000, twice the murder rate of the next-closest MSA. We have decided to remove New Orleans, believing that this case requires its own investigation. Here are scatter plots of some of the most explanatory features versus the murder rate:



Overall, we found that family structures that are 'less traditional' (e.g., unmarried households, unmarried mothers with babies, households where husband is missing) have a positive relationship with murder rate. Additionally, lack of education and income also correlate positively with murder rate.





3. A revised project question based on the insights you gained through EDA.

a. We intend to measure to what degree the underlying features that we have measured are explanatory of murder rates. To get a quantitative measure of this, we plan to create three models:
   i. MSA + year; by one-hot encoding each MSA we get bucket variables that represent all of the underlying factors that differ between MSAs- both those represented in our dataset and those that are not. The year allows us to capture overall trends through time.
   ii. MSA + year + features; we introduce our selection of features into the above model to observe how this affects model performance and reliance on the MSA encoded variables.
   iii. Features + year; We remove the MSA encoded variables, and evaluate model performance based only on our selected features.

b. Through a comparison of these three models, we expect to gain insight into the existence or non-existence of significant underlying causes of violent crime which are not being collected with census data.