Introduction

In this study, we analyze crime data from various cities and states in the USA to identify significant predictors of different types of crimes and understand the underlying patterns and trends. By applying concepts learned from our coursework, we aim to build models that accurately represent the relationships between socio-economic factors and crime rates, specifically focusing on how these factors influence the housing price index. A significant part of our analysis is examining how population size and violent crime rates can be used to predict changes in the housing price index over multiple years for the same cities. Understanding the relationship between predictors like population size, violent crime rates, and the housing price index will provide a broader context for the data. This analysis will highlight the socio-economic factors that contribute to the fluctuations in the housing prices and how these factors interact with crime rates, understanding the influence of population changes on housing prices can help government agencies to develop new initiatives to manage urban growth, housing affordability and crime rate.

Exploratory Data Analysis

Correlations

```
# Method 1: Use a corrgram
corrgram(crimeData,upper.panel=panel.cor)
# Method 2: Use a correlation plot
cor_data <- crimeData %>% select(HPI,Population, TotalCrime,Year)
cor_matrix <- cor(cor_data)
corPlot(cor_matrix,cex =0.6)
```

From the corrgram, we see that each individual crime (Rape, Robberies, Assaults, and Homicides) is all heavily correlated to one another (see exhibit 1). They are also all heavily correlated with total crime, which is intuitive as they directly impact the total number of observed crimes. Moreover, we see that each crime (and total crime) is correlated with the population size. Which makes sense as we expect to see a rise in crime with a bigger population. The final correlation observed is between year and HPI. HPI does vary year to year (as economic conditions vary year over year influencing housing prices), so this relationship is expected. Our correlation plot further solidifies the results seen from the corrgram, see exhibit 2.

Interactions between a continuous and a categorical predictor

Our data had two categorical variables, city and state. Thus, we created 5 interaction plots to measure the relationship of continuous variables within different states and the cities within those states. When comparing the HPI by year in each state and city, we see fairly similar results across the American states, although we can observe distinct variances between cities in the same state. When looking at population by year we see similar trends across the states, but there are clear variances between certain cities (even ones in the same state; but obviously some metropolises will have larger populations). For total crime by year, we once again can observe a specific trend across most states. States such as Illinois, Texas, Michigan and Pennsylvania have noticeably different trends (but this looks like it's being driven from cities such as Detroit, Chicago, Houston and Philadelphia - cities more prolific for crime). When comparing HPI by population size it's hard to observe any distinct trends in both American states and cities. For total crime vs population size, it appears that generally for most states a smaller population

equates to a smaller number of total crimes, but within certain cities (namely, Chicago) this trend doesn't necessarily tie out (see exhibit 3).

We also created a scatter plot to visualize the relationships between HPI and Year, HPI and total crime. It was clear that each relationship was non-linear, especially the relationship between HPI and Year (see exhibit 4)

Fitting a model by hand and performing an ESS test

```
dummy_variable1=model.matrix(~City, data = crimeData)
dummy_variable1=dummy_variable1[,-1]
dummy_variable2=model.matrix(~State, data = crimeData)
dummy_variable2=dummy_variable2[,-1]

y3<-crimeData$HPI
A3<-model.matrix(~ Year + TotalCrime + City + City:TotalCrime, data =crimeData)

B3=solve(t(A3) %*% A3) %*% t(A3) %*% y3
SSE3=t(y3)%*%y3-t(B3)%*%t(A3)%*%y3
SSR3=t(B3)%*%t(A3)%*%y3-nrow(A3)%*%(mean(y3))^2
SST3=t(y3)%*%y3-nrow(A3)%*%(mean(y3))^2
r_squared3=SSR3/SST3
```

When trying to build various models by hand, we ran into several errors and limitations. Firstly, including both categorical variables (state and city) within the same linear model, it had resulted in too many aliased variables. Moreover, including Year and Population within the same linear model created problems as well. For example, when we performed a VIF test for a model that included both categories (as well as total crime and city), we got a VIF of ~29 for population, ~20 for Chicago:Population and ~5 for total crime, suggesting serious multicollinearity. We also couldn't create a model that included all 4 individual crimes with total crime (which made sense due to the observed correlation from the correlation from the corrgram). Finally, the addition of the following interaction terms created singularity issues when solving the model by hand: Year:Population, Year:TotalCrime, Population:TotalCrime, Population:State.

For the models we were able to test (Robberies+Assaults+Homicides+Rapes+Population+City, TotalCrime+Population+State, TotalCrime+Year+State, Year+TotalCrime+City+City:TotalCrime to name a few), we observed some promising results. Models that included Year yielded high R^2 values, whereas models with population resulted in some of the lowest observed r^2 values. This is likely due to the correlation Year had with HPI. It also seemed that the inclusion of the city resulted in better models than the ones with State.

From the information above, we can determine what combination of predictors will create the "best" linear model. Firstly, since we cannot include both categorical variables (due to the observed errors), we should choose city over state as it had yielded better results. Additionally, we cannot use both population and year together due to the issues we previously encountered and thus, we will choose year over population based on our findings above. It's also trivial to choose Total Crime over each individual crime as it encompasses the information (less complexity in the model). Individual crimes act more as proxies for total crime. Finally, we will include interaction terms in our final model where it is possible to improve the accuracy of the model (Especially between City and Total Crime based off the interactions we identified earlier from our plots).

Investigate transformations of the Predictors and of the response

```
lm1<-lm(HPI ~ Year + TotalCrime + City + City:TotalCrime, data =crimeData)
summary1<-summary(lm1)
summary1$adj.r.squared
stepwise_approach<-stepAIC(lm1, direction="both")
## Start:  AIC=11105.34
summary<-summary(stepwise_approach)

lm2<-lm(sqrt(HPI) ~ TotalCrime + Year + City + City:TotalCrime, data =crimeData)
summary2<-summary(lm2)
lm3<-lm(log(HPI) ~ Year + TotalCrime + City + City:TotalCrime, data =crimeData)
summary3<-summary(lm3)
lm4<-lm(log(HPI) ~ TotalCrime + poly(Year, 3, raw = FALSE) + City + City:TotalCrime,
data =crimeData)
summary4<-summary(lm4)

BC=boxcox(lm1)
L=BC$x[which.max(BC$y)]
crimeData$HPI_BCM=((crimeData$HPI)^L-1)/L
lm5_BCM=lm(HPI_BCM~Year + TotalCrime + City + City:TotalCrime, data = crimeData)
summary5<-summary(lm5_BCM)
summary5$adj.r.squared
## [1] 0.8882729
stepwise_approach1<-stepAIC(lm4, direction="both")
## Start:  AIC=-229.78
## [1] 0.8922871
```

Out of all the linear models we investigated, we chose the model with the greatest r^2 to investigate transformations to further improve its fit and stabilize the variance. We first applied a stepwise AIC approach to see if there are any modifications we needed to make to the model. Interestingly, we observed that the AIC's multiple R^2 value was the same as the one we manually calculated (indicating the model did not change much), the adjusted R^2 had the same result too (0.8388072). It also had a very high AIC indicating it was not fitting the data well (although it did determine total crime: city & year were both important predictors). We then decided to apply some transformations to improve the model's fit. We then checked the initial model's diagnostic plots to see what transformations we should apply (see exhibit 5). The QQ-plot highlighted that the residuals were not normal as tails trailed off the 45-degree line. From the residuals vs leverage we do see some influential points (this could be Chicago as it has caused issues earlier), making us question if we should consider removing outliers. From earlier scatter plots, we saw a non-linear relationship between year and HPI (see exhibit) implying we should add a polynomial transformation. The scale location is a curved line oppose to a straight horizontal one. Thus, we determined that we could attempt a log transformation or a sqrt transformation to help with the observed issues.

We began by applying a basic sqrt transformation, which led to improvements in both the diagnostic plots and the adj. r^2 value (0.8789059). The QQ-plot of the initial model determined skewness and thus, we decided to also try a log transformation to improve our results. From both the diagnostic plots and the adj. r^2 value (0.8867949), we again saw improvements. We decided to expand our transformations by adding an orthogonal polynomial to year (to prevent correlation between polynomial terms) due to its non-linear relationship with HPI. Once again, we saw an improvement as the adjusted R^2 value became 0.9167559 and the diagnostic plots

had improved (see exhibit 6). Finally, we applied a box-cot to just see if we can further improve our results (as theory states we can move on to it if Log & sqrt not satisfactory). Surprisingly, it did not result in the highest r^2 value (0.8882729). Based off the diagnostic plots, we determined that the fourth linear model (lm4) was the most superior. We verified this assumption by comparing its AIC to the AIC of the original model. This model had a significantly lower AIC indicating it's a better fit.

Model Fitting

As discussed above, when looking at the diagnostic plots between our initial linear model and its transformed version, we saw big improvements. Please see exhibits 5 and 6. The summary of the transformed version also highlighted that based on the p-values, each polynomial term for year was significant, some of the interaction terms for total crime and city were significant (as were individual cities without the interaction term) and total crime was not significant (which was surprising, especially as some total crime: city predictors were significant). Based on the high R^2 value, the promising diagnostic plots and the combination of predictor's significance in predicting HPI, we chose linear model 4 as our final model.

Permutation Test for Overall Significance

```
set.seed(123)  # For reproducibility
n_permutations <- 1000  # Number of permutations
perm_f_stats <- numeric(n_permutations)
for (i in 1:n_permutations) {
  permuted_data <- crimeData
  permuted_data$HPI <- sample(permuted_data$HPI)  # Permute the response variable
  perm_model <- lm(HPI ~ Year + TotalCrime + City + City:TotalCrime, data =
permuted_data)
  perm_f_stats[i] <- summary(perm_model)$fstatistic[1]
}
p_value <- mean(perm_f_stats >= original_f_stat)
cat("Original F-statistic:", original_f_stat, "\n")
```

```
## Original F-statistic: 101.9409
cat("Permutation p-value:", p_value, "\n")
## Permutation p-value: 0
```

We designed the permutation test to randomly shuffle the Housing Price Index (HPI) values 1,000 times. Each iteration recalculated the F-statistic and from there, we were able to compare the distribution of F-statistics to the original model's F-statistic. Furthermore, the test resulted in a p-value that was extremely close to zero, which indicated that the model we created is significant. We can conclude there is a distinct relationship between the HPI and the chosen predictors (Year, Total Crime and City).

Compare Observed Results to Simulated Results

```
original_coef <- coef(lm4)
original_r_squared <- summary(lm4)$adj.r.squared
residual_variance <- var(resid(lm4))

n_simulations <- 1000  # Number of simulations
simulated_r_squared <- numeric(n_simulations)
set.seed(123)
```

```
for (i in 1:n_simulations) {
  simulated_response <- predict(lm4) + rnorm(n = nrow(crimeData), mean = 0, sd =
sqrt(residual_variance))
  sim_model <- lm(simulated_response ~ TotalCrime + poly(Year, 3, raw = FALSE) + City
+ City:TotalCrime, data = crimeData)
  simulated_r_squared[i] <- summary(sim_model)$adj.r.squared
}
simulated_r_squared_df <- data.frame(R_squared = simulated_r_squared)
```

Despite simulating new housing prices over 1000 different simulations, our model's r-squared value did not significantly change and was always over 0.90. As shown in Exhibit 7, the distribution of R-squared values did not vary greatly as it yielded results consistently between 0.91-0.93. Additionally, we see that our initial model's adjusted R^2 value laid in the range where most of the simulated R^2 values were distributed. Despite random variations, our model performed consistency well, speaking to how strong the model's explanatory power is.

Limitations:

Firstly, the simulated response variable assumes that the residuals follow a normal distribution with constant variance, which may not hold true in the actual data. If the model's assumptions are not held together, the results of the simulation would not be accurate. Additionally, the model was simulated 1,000 times, which may not fully capture all possible variations. Finally, the analysis focuses on the adjusted R-squared metric, which does not emphasize other important model characteristics like prediction accuracy or the significance of individual predictors. The data set under analysis can fluctuate caused by various economic factors such as recessions, policy changes, or market instability. These economic fluctuations can lead to significant variability in the data, which the model may struggle to capture accurately. It is also important to note that since we're dealing with data across multiple years, the data is no longer independent which creates large limitations such as autocorrelation and inflated R^2 values.

Conclusion:

The simulation of the adjusted R-squared values based on the final linear model (lm4) provides insight into the accuracy and reliability of the model's fit. The histogram of the simulated R-squared values reveals how much variability in the adjusted R-squared can be expected due to random noise in the data. By comparing the original adjusted R-squared value to the distribution of simulated values, we can assess whether the original model's fit is unusually high or within a typical range. The R-squared value of the model is 0.9167, it indicates that the model explains approximately 91.67% of the variance in the dependent variable, suggesting a strong fit. This high R-squared value implies that the model captures most of the variability in the data, making it a reliable tool for prediction within the context of the dataset. The parameters for total crime and the polynomial terms for year were all negative and around the same value (expect the cubic term which has the biggest impact on HPI). This makes sense as crime would likely bring down housing prices. It is interesting that an increase in the year would have a negative impact on housing prices. Moreover, the parameter for total crime : city varied between positive and negative indicating cities with more crime (such as Chicago or Columbus) bring down prices, whereas cities (ex: Fresno) with less crime increase housing prices.

Executive Summary:

The goal of this study was to create a model that accurately articulates the relationship between housing prices and socio-economic factors including population and crime within the United States. We had hypothesized that the aforementioned socio-economic factors, namely crime, directly influences the housing price index within the United States and wanted to bring this claim to fruition.

We began by measuring the correlations between the socio-economic factors we had chosen for our experiment to identify the risk of dependencies. We found that the factors year and HPI, as well as crime and population size were correlated. We then further examined the interactions between each factor to have a better sense of their relationships to one another.

After thoroughly understanding the relations between our variables we began testing models with numerous combinations of predictors (i.e. variables). Due to the high multicollinearity between many of the variables, we were limited in our choice in the combinations of predictors. For example, we observed difficulties in models that included both year and population.

Eventually, we determined the model that yielded the best results and could best explain the variance found inside the model. The model encapsulated year, total crime, cities, and the interaction between city and total crime to predict the housing price index. Not only did these predictors help create the most effective model, we believe that they are also the most meaningful in terms of context. Including year allows us to consider housing prices year after year. The inclusion of total crime captures the impact of crime on housing prices. Ensuring that our data includes the city, keeps in mind the variation of housing prices in different locations. Finally, the city and total crime interaction considers the variation of how crime impacts HPI in cities within the US.

Despite the excellent results the model boasted, we further improved its accuracy by applying a series of transformations to both the predictors and HPI. Transformations are a mathematical approach that improves the fit of a model. We deliberately applied our transformations to improve our model in the areas in need of improvement. We were successful in our efforts as the final transformed model achieved an adjusted R-squared value of 0.9168. In layman's terms, our model can explain ~92% of the variability in housing prices. Even after 1000 simulations that introduced random variations, our model was still able to produce R-squared values between 0.91-0.93 and a p-value extremely close to 0 (The closer a p-value is to zero the greater the significance), indicating our model's significance. The p-values for the individual predictors determined the significance of year, cities and crime within cities. Allowing us to conclude these variables are crucial in predicting house price within the United States. Overall, our model determined that crime and year had a negative impact on housing prices, whereas the city and crime within a city could positively or negatively impact housing prices depending on the location.

It's important to highlight that even with our promising results, there are countless other factors and trends that affect housing prices that our model failed to consider.

Exhibits:

Exhibit 1: Corrgram

Exhibit 2: Correlation Plot

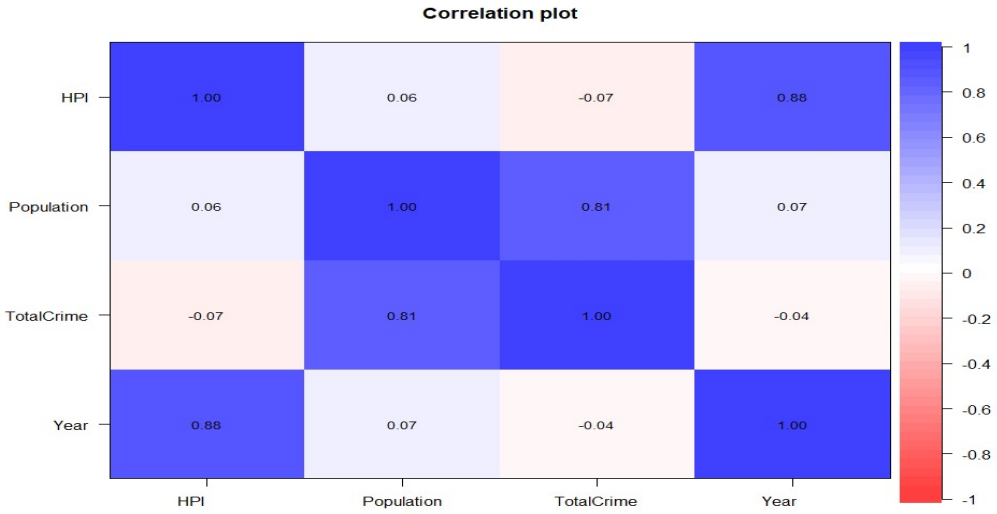

Exhibit 3: Scatterplot between Year vs HPI

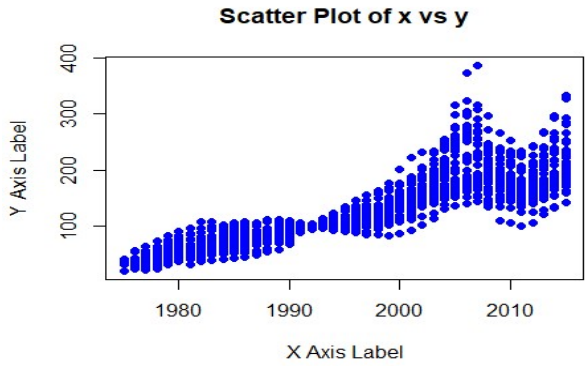

Exhibit 4: Interaction Plot Between Pop., Total Crime, City and State

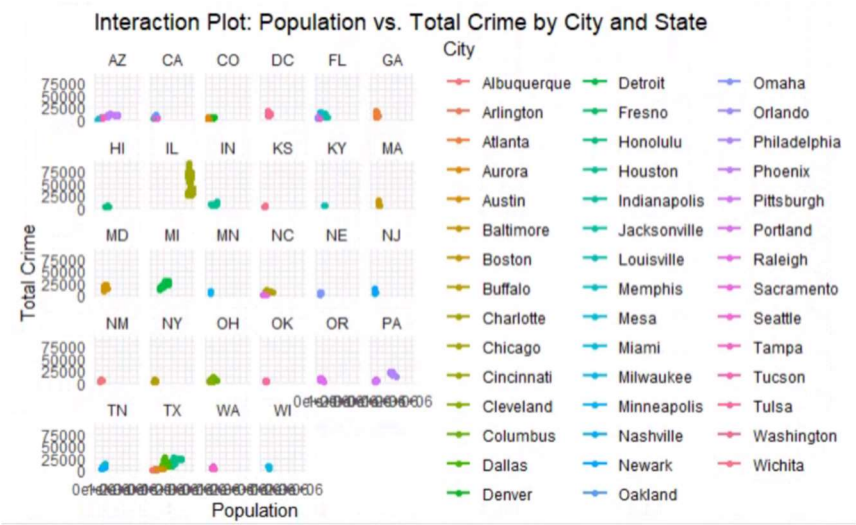

Exhibit 5: Diagnostic Plots for Linear Model 1

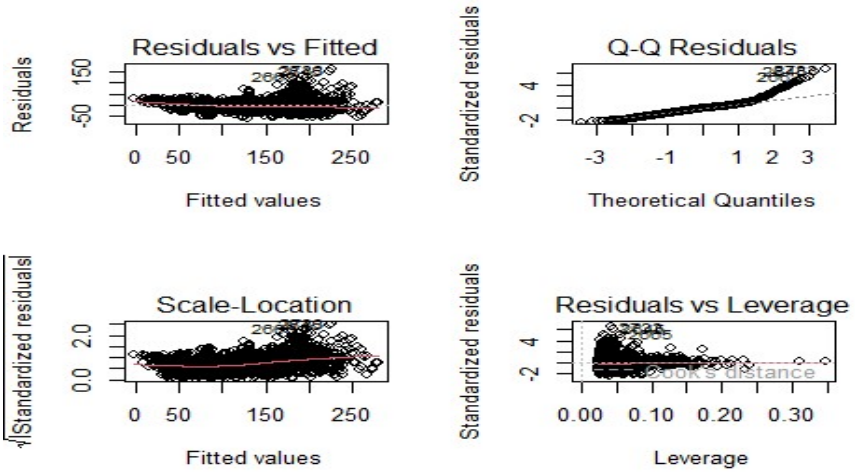

Exhibit 6: Diagnostic Plots for Transformed Version of Linear Model 1

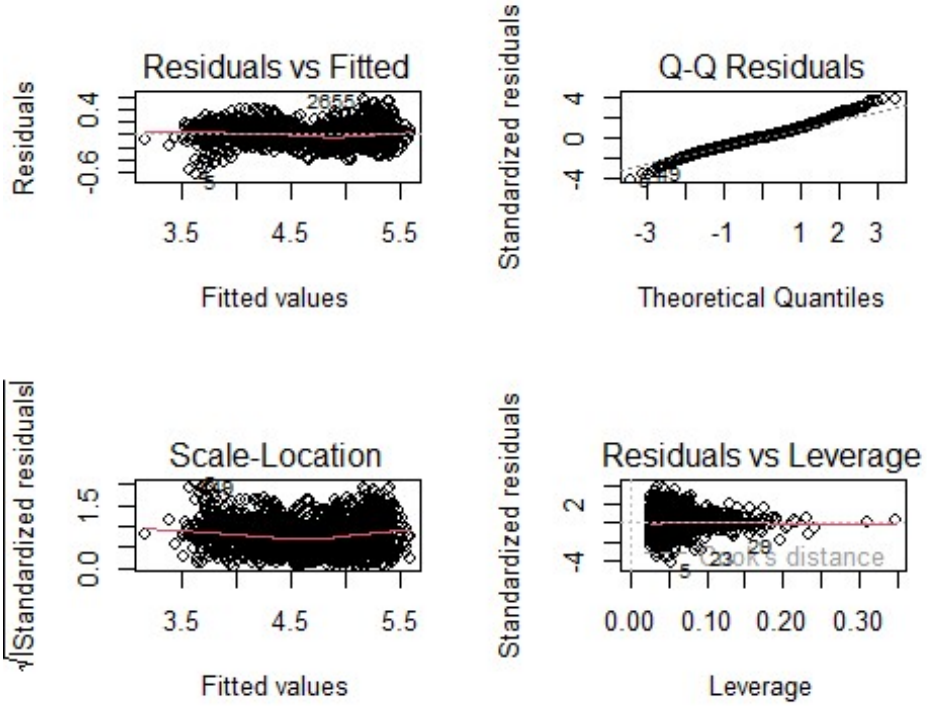

Exhibit 7: Simulated Adjusted R-squared values