Springboard

Data Science Cohort 2024

Capstone Project

Ofri Oren

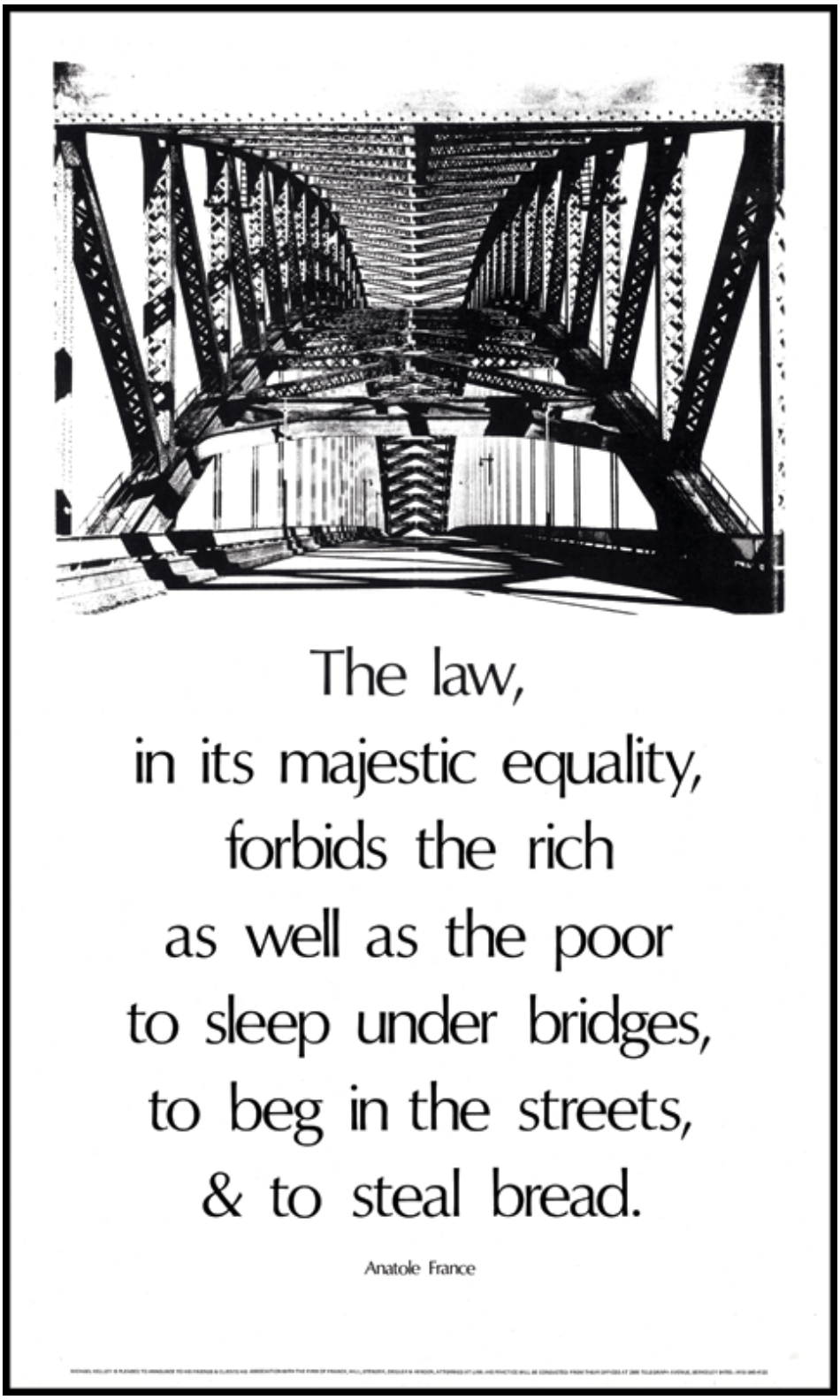
Homelessness in California

Summary:

I started with an exploration of the potential correlates of homelessness in California but could not pinpoint any of the factors as directly or strongly causal. I thus built a model using data from previous years to incorporate temporal trends into a forecasting model to predict homelessness in California by county. I tested different models but found the vector auto-regressive model had the best predictive power based on Akaike information criteria (AIC) and Bayesian information criteria (BIC) scores. Although these data do not provide any prescriptive measures to address the root cause of homelessness, they can allow people to alleviate the problem by preparing shelters and resources for the predicted amount of homeless in their county.

Outline:

- Background
- Data
- Exploratory Data Analysis
- Processing and Modelling
- Forecasting and Results
- Bibliography and Datasets



Artist: David Lance Goines; Anatole France Poster; [https://www.goines.net/Gallery/gal_xtra/007_anatole_france.gif] accessed on 2 October 2024.

Background

In a recent Supreme Court ruling on June 28, 2024, the court decided that police have the right to cite and fine people who camp or sleep in public spaces. The criminalization of homelessness is an ongoing process in California, including measures like those in San Francisco that prohibit sitting or lying down on public sidewalks. However, with limited shelter space available many people have no alternatives. A possible response could be to provide sufficient resources for people to stay in shelters and eventually return to housing.[1] While my initial interest in this project was to look for correlation between healthcare cost and availability and lack of housing, I eventually turned towards building a model that could predict levels of homelessness per county in California to estimate the scope of the problem in upcoming years. In this project I look at some of the potential correlates of homelessness and build a model to forecast homelessness based on statewide economic factors and previous homelessness counts per county.

Data

Variables included in the final dataset:

Healthcare cost, California: aggregate spending for all payers in the state, measured in millions of dollars.[2]

[1] Ludden, Jennifer. 06/28/2024. "The Supreme Court says cities can punish people for sleeping in public places." *Morning Edition.* National Public Radio Broadcast, accessed at [https://www.npr.org/2024/06/28/nx-s1-4992010/supreme-court-homeless-punish-sleeping-encampments] in September 2024.
[2] Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group. "Table 1: Total All Payers State Estimates by State of Residence (1991 - 2020) - Personal Health Care (Millions of Dollars)." [https://www.chcf.org/publication/2023-edition-california-health-care-spending/#related-links-and-downloads]

Housing prices, California: All-Transactions House Price Index for California [CASTHPI]. It is the annual average house price index increase where first quarter equals one hundred, e.g. an amount of 200 indicates a twofold increase in prices.[3]

CPI, urban consumers: the cost price index for urban consumers across California.[4]

CA unemployment rate: the unemployment rate in California.[5]

Obama ACA indicator: categorical variable that indicates times prior to the Affordable Care Act as zero, passage of ACA through legislature in 2010 with universal mandate as one, and passage of the 2018 legislature revoking the universal mandate as two.[6]

Overall homelessness, California: based on the homelessness by state dataset, actual numbers.[7]

CA-500 through CA-614: based on the homelessness by county dataset, per California county, actual numbers.[8]

CA-500	San Jose/Santa Clara City & County
CA-501	San Francisco
CA-502	Oakland, Berkeley/Alameda County
CA-503	Sacramento City & County

[3] U.S. Federal Housing Finance Agency, All-Transactions House Price Index for California [CASTHPI]. Retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CASTHPI, June 2, 2024.
[4] California Government dataset, https://www.dir.ca.gov/OPRL/CPI/CPICalculator/CpiCalculator.aspx, accessed 2024.
[5] U.S. Bureau of Labor Statistics, Unemployment Rate in California [CAUR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CAUR, June 2, 2024.
[6] [https://en.wikipedia.org/wiki/Affordable_Care_Act], accessed in 2024.
[7] [https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/], accessed 2024.
[8] [https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/], accessed 2024.

CA-504	Santa Rosa, Petaluma/Sonoma County
CA-505	Richmond/Contra Costa County
CA-506	Salinas/Monterey, San Benito Counties
CA-507	Marin County
CA-508	Watsonville/Santa Cruz City & County
CA-509	Mendocino County
CA-510	Turlock, Modesto/Stanislaus County
CA-511	Stockton/San Joaquin County
CA-512	Daly/San Mateo County
CA-513	Visalia/Kings, Tulare Counties
CA-514	Fresno City & County/Madera County
CA-515	Roseville, Rocklin/Placer County
CA-516	Redding/Shasta, Siskiyou, Lassen, Plumas, Del Norte, Modoc, Sierra Counties
CA-517	Napa City & County
CA-518	Vallejo/Solano County
CA-519	Chico, Paradise/Butte County
CA-520	Merced City & County
CA-521	Davis, Woodland/Yolo County
CA-522	Humboldt County
CA-523	Colusa, Glenn, Trinity Counties
CA-524	Yuba City & County/Sutter County

CA-525	El Dorado County
CA-526	Amador, Calaveras, Mariposa, Tuolumne Counties
CA-527	Tehama County
CA-529	Lake County
CA-530	Alpine, Inyo, Mono Counties
CA-531	Nevada County
CA-600	Los Angeles City & County
CA-601	San Diego City and County
CA-602	Santa Ana, Anaheim/Orange County
CA-603	Santa Maria/Santa Barbara County
CA-604	Bakersfield/Kern County
CA-606	Long Beach
CA-607	Pasadena
CA-608	Riverside City & County
CA-609	San Bernardino City & County
CA-611	Oxnard, San Buenaventura/Ventura County
CA-612	Glendale
CA-613	Imperial County
CA-614	San Luis Obispo County

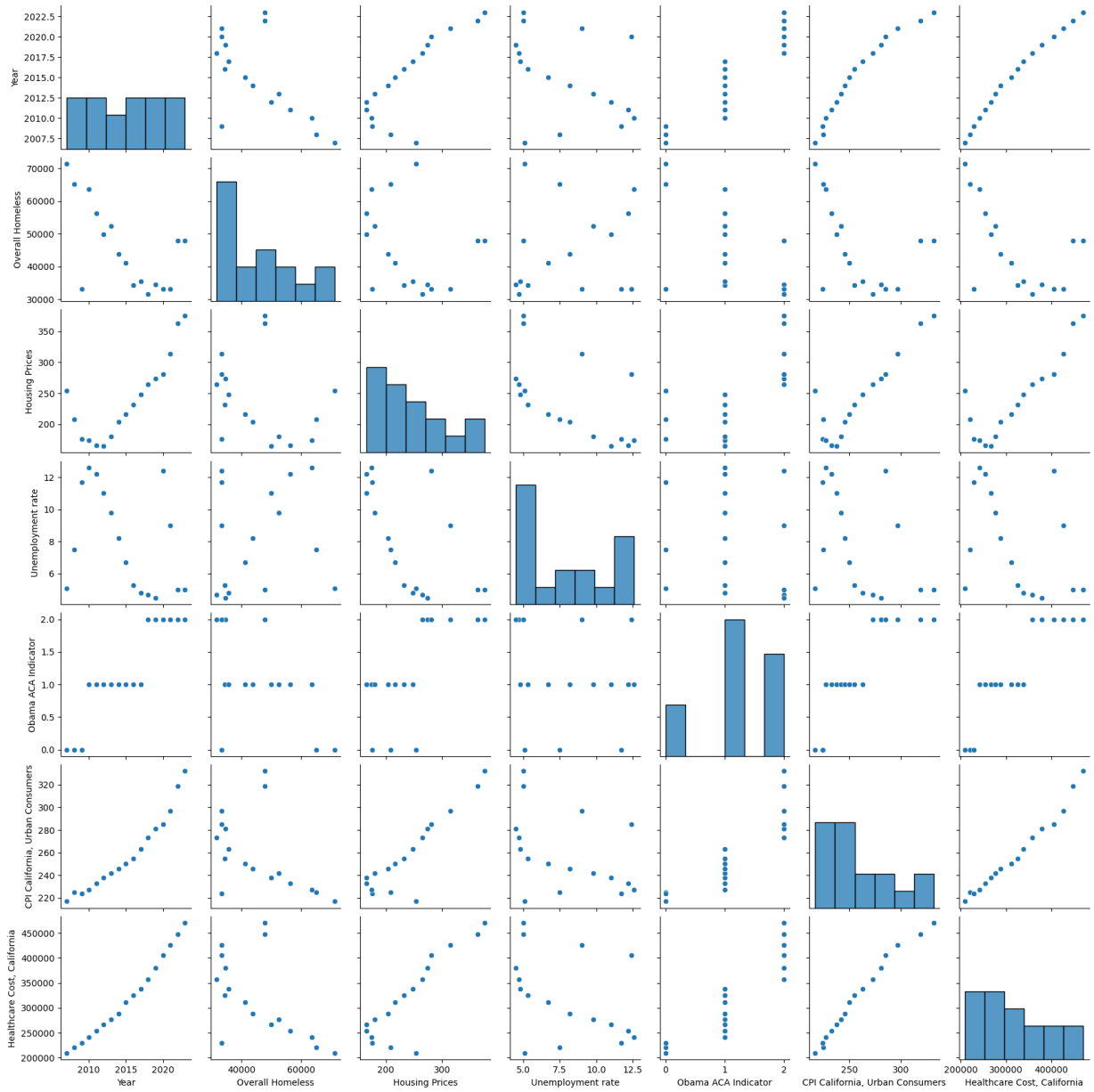
Data wrangling and cleaning

I initially focused just on California and wrangled data by combining multiple datasets into one data frame. I found information for the years 2007 – 2023 for all the variables except healthcare costs in California. I found the yearly rate of change from year to year and noticed that the percent change was dynamic, so I took the average of the years present and predicted it would increase by the mean rate of increase for the missing year, 2023. I later added in homeless data from counties across California instead of just Los Angeles and had missing values for some of the counties. I averaged the amount of homelessness *per county* and used those values to fill in my missing data for that dataset. Instead of joining data, I often just concatenated the columns or rows required for my inquiry. I created the categorical variable for the Affordable Care Act and only had numerical data apart from this, so I did not need to create dummy variables.

Exploratory Data Analysis

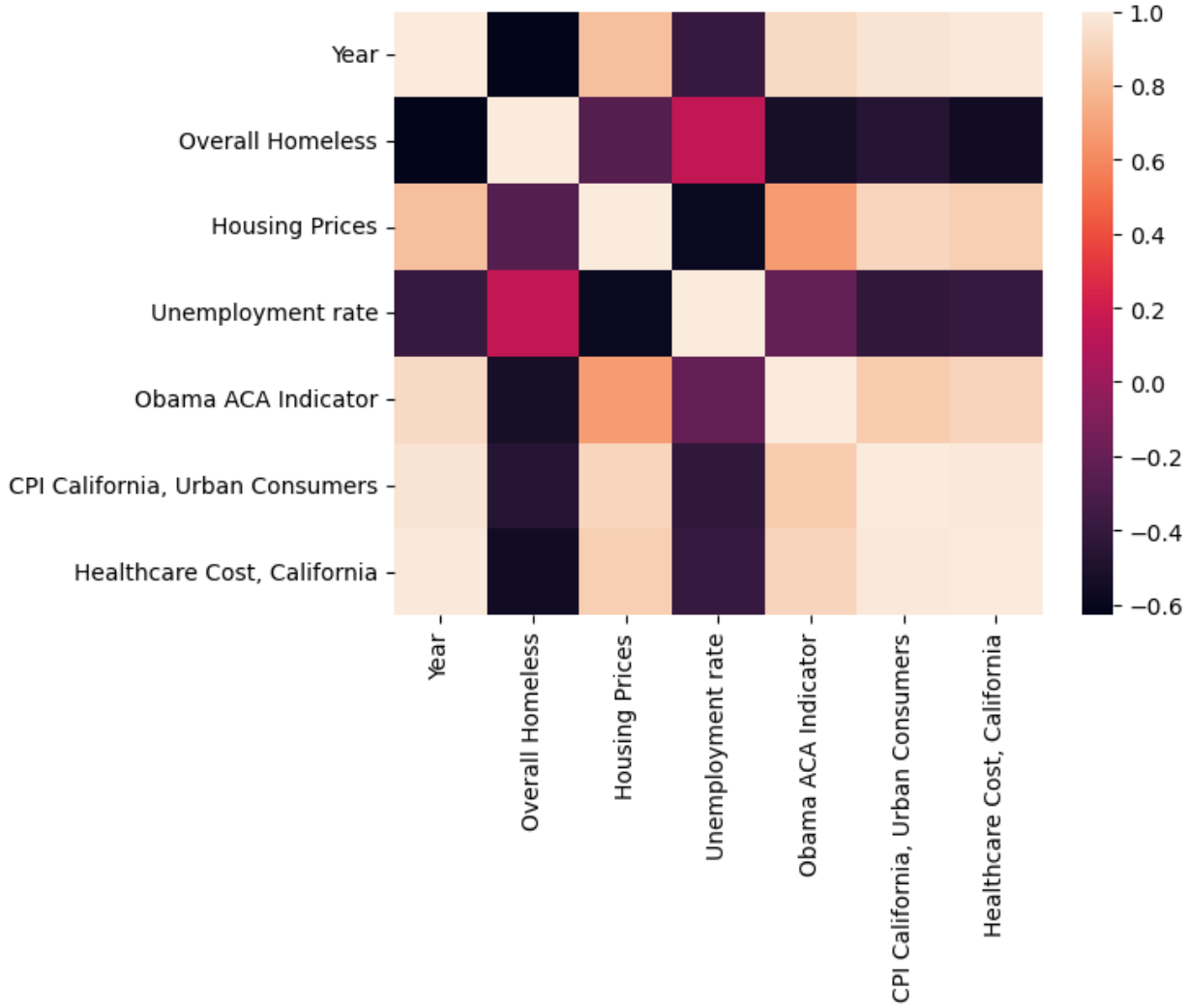
I began with the following dataset looking at Los Angeles specifically and the effect of healthcare costs on homelessness. I first did a pair plot to look for correlation between variables, normality in distribution, and general trends in the data. Much of the data shows a strong linear correlation, however some of it is positive and some of it is negative, so it is not immediately clear what the general trend for homelessness in California would be. Initially, I just looked at the indicator variables without the per-county data.

Pair-plot of Indicator Variables



Next, I made a heatmap of the correlation between the variables to do a quick assessment of positive or negative correlation and the strength of the correlation. In the first heatmap, there is only the indicator variables apart from the per county data.

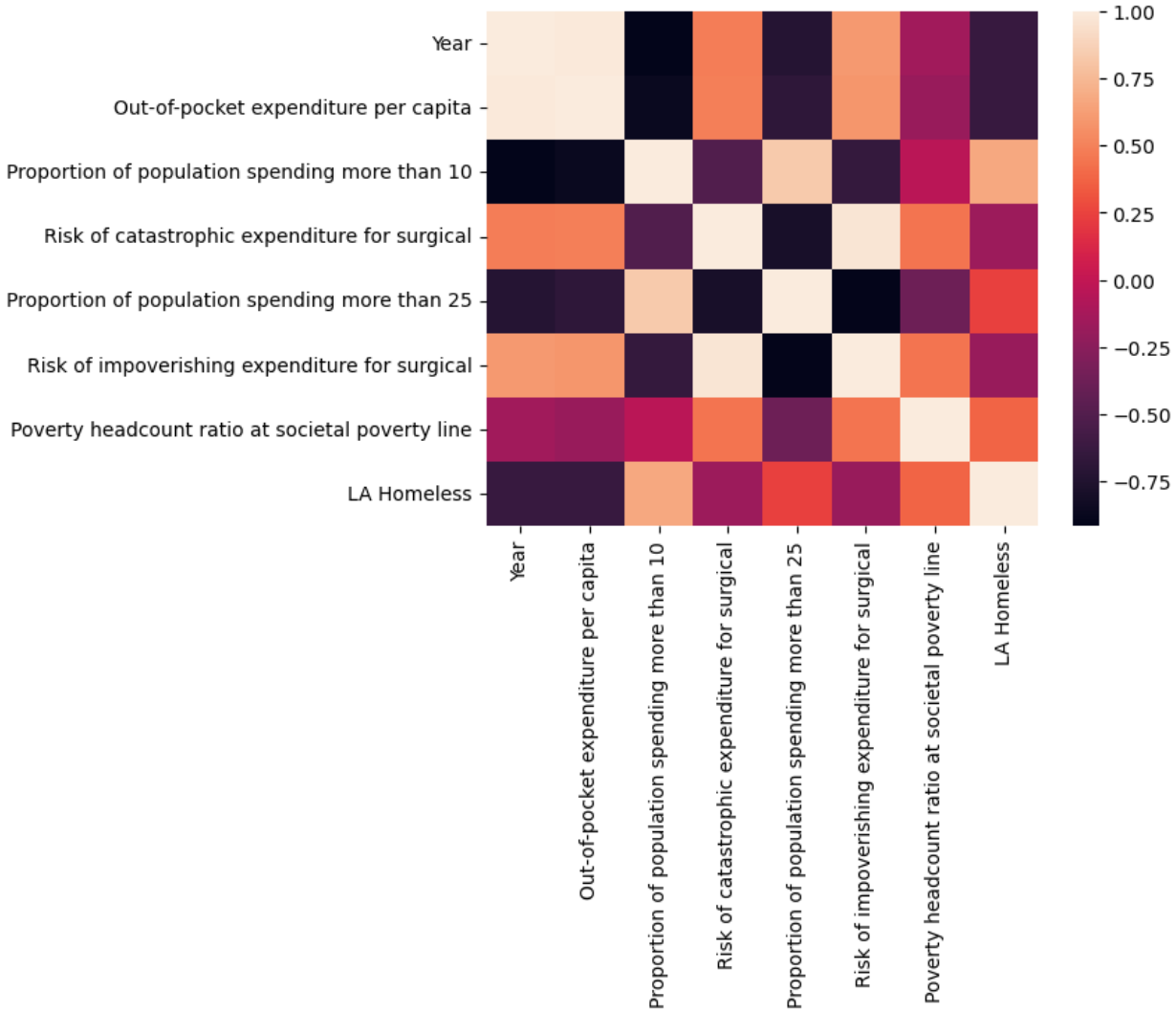
Heatmap of Correlation Between Variables



Initially, I also gathered data from the World Bank datasets on percentage of population spending more than 25% of income on out-of-pocket healthcare costs, risks of impoverishing surgical or other health expenditures and so on. However, the data from these datasets is aggregated from the US as a whole and might not reflect changes in Los Angeles specifically. I decided to omit it and did not reintroduce it for the California-wide county evaluations as a predictor. The distances and variety between states made it seem to have lower predictive power

to include country-wide data for predictions in California rather than state-wide data for Los Angeles or California as a whole.[9]

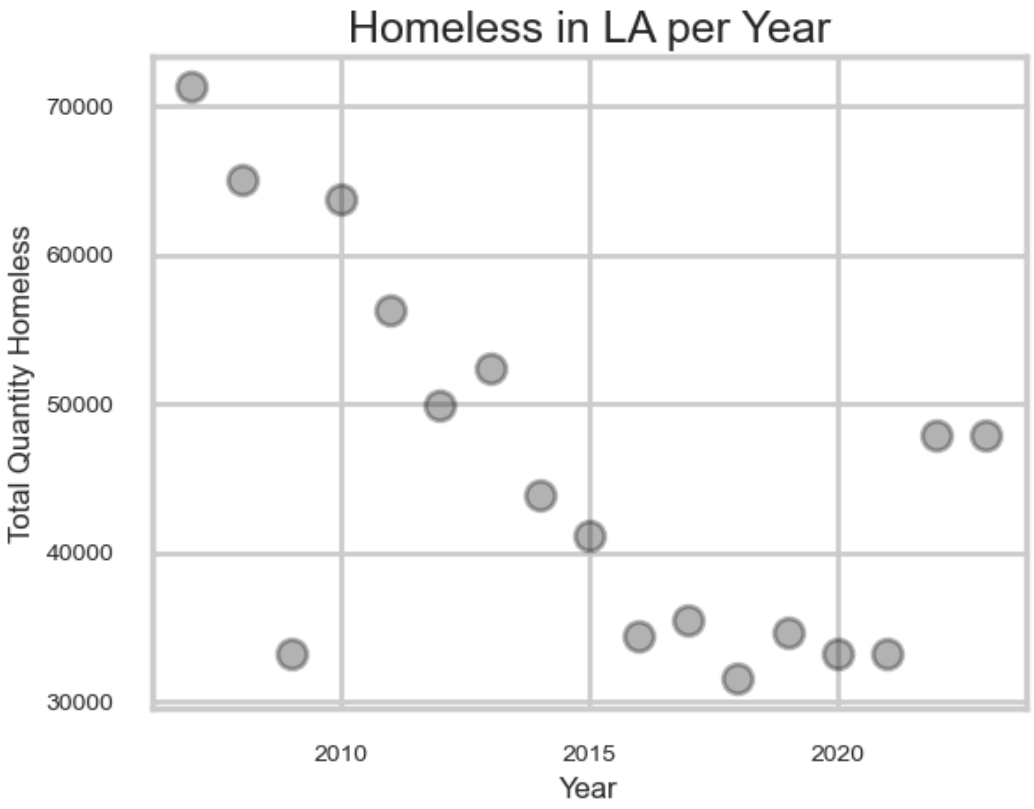
Heatmap of Correlation Between World Bank Indicators

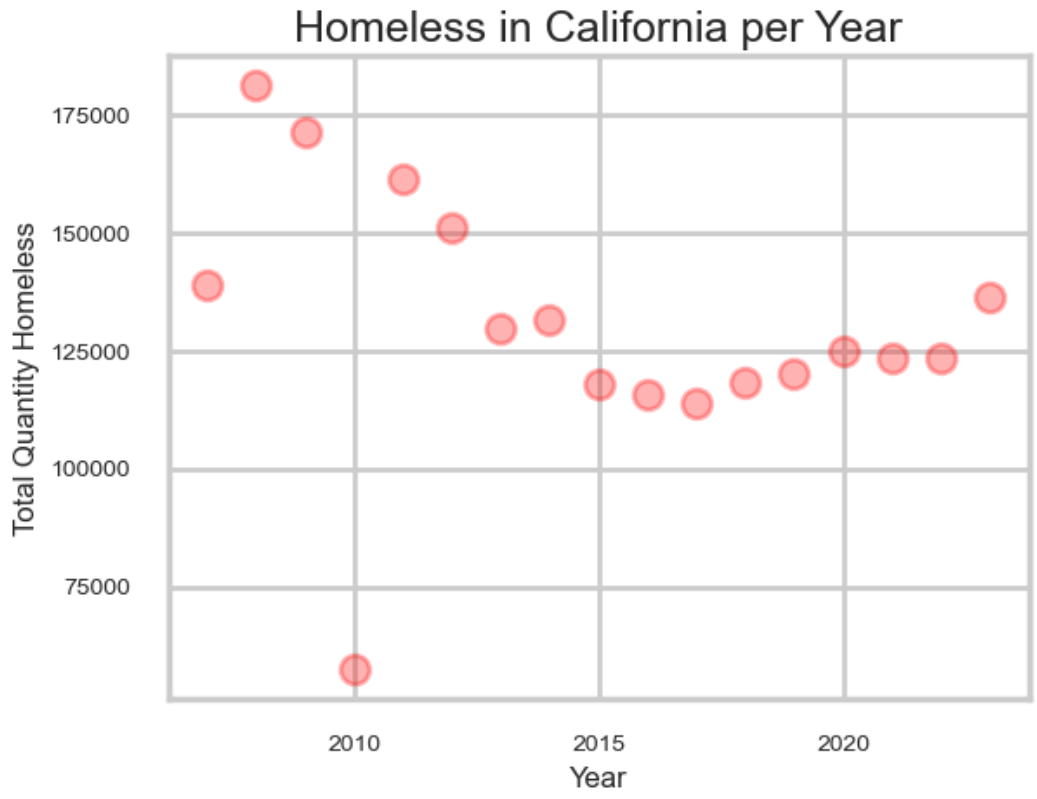


As indicated by the darker squares for year, much of this data is strongly correlated with time. This indicates that time series modelling would be a good fit for this data. However, I first

- [9] The World Bank, World Development Indicators (2024). *US Health Indicators; Poverty and health indicators*. Retrieved from [https://databank.worldbank.org/source/world-development-indicators] in 2024.

looked at the overall trends in the data for Los Angeles and then California.





Processing and Modelling

Both graphs show similar trends, a decrease in homelessness until about 2020 and then an upwards trend. I tried to find the line of best fit using a linear model, an exponential model, and then completely shifted to using time series models. I scaled the data to ensure that the indicators like price indices, housing costs, healthcare costs, and others would be in the same frame of reference as the homeless population measures. I decided to try the auto-regressive integrated moving average (ARIMA) model first to make predictions, but this model does not do well with too many variables. Thus, I used a principal component analysis (PCA) to isolate the principal indicators of fluctuation in the data. I checked to see what percent of the variation in the data was accounted for with each principal component and found that one principal component accounted

for almost all the variation. However, I chose to use two components to account for additional variation but not all the variation to allow for some error and to prevent over-fitting. It reduced the error for the predictions of past data to minimize the error for 'future' predictions of 2023, where there is a shift from the general trend. In the second vector auto-regression (VaR) model, there is no decrease in power of the test due to an increased number of variables, so I did not have to use the PCA reduction.

I then split the data into training and test sets using the time series specific split. This splitting function in *sklearn* makes sure that the data is split without shuffling, i.e. not randomly, so that there are no backwards predictions. I ran a simple linear model and generated mean-squared errors (MSE) and r-squared (R^2) values to have a basis for comparison. I plotted the auto-correlation function (ACF) for each feature in the indicator set of variables to see how many lags to use for the time series. I ran an AD Fuller test to ensure stationarity and to make sure that the correct number of lags was three, as shown for almost all variables. I first ran a simple ARIMA model with one lag. I then ran an ARIMA model with three lags to match the peak shown in the ACF plots. I collected MSE and the average MSE for comparison. The second ARIMA model showed substantial improvement from the first, shown by the decrease in MSE, and I decided to use three lags in the VaR model. I ran the VaR model to make predictions and saved the MSE as well as AIC and BIC scores as indicators of model performance.

Forecasting and Results

The result metrics for the models show that the first two models (ARIMA1 and ARIMA2) have lower residuals, as it also appears in the MSE-scores. While the MSE-score appears much higher for the VaR predictor the AIC and BIC scores are quite low for the VaR

model. The low AIC and BIC scores indicate that overall, the VaR model is a better predictor for this data. Below is a table of the metrics measuring the fit of the model to the data and the accuracy of the predictions, it is grouped by type of metrics with MSE first and AIC/BIC second.

Table of Metrics per Model

Model	**Metric 1**	**Metric 2**
Linear Regression (without PCA)	MSE : 795024.4536907848	R^2: 0.8084225410973096
ARIMA 1 (one lag)	MSE Scores: [0.8232445008828726, 0.28875381936427336, 0.16689813981394258, 0.4284484050792737, 0.0767696582109092]	Average MSE: 0.3568229046702543
ARIMA 2 (three lags)	ARIMA2 MSE Scores: [1.925674247900199, 0.2028152275174579, 0.09855641689779224, 0.42116319640969735, 0.05163774483800012]	Average MSE: 0.5399693667126293
VaR	VAR MSE Scores: [354806659460.1925, 3258731.234438983, 235845996.07972065, 11479199.778941542]	Average MSE: 88764310846.8214
ARIMA 1 (one lag)	AIC: 296.384	BIC: 302.238
ARIMA 2 (three lags)	AIC: 137.651	BIC: 143.003
VaR	AIC: -144.141	BIC: -132.773

Based solely on metrics, VaR seems to be our best model by far. However, these results can be misleading without looking at the actual data behind them. The predicted results from the VaR model for overall homelessness in 2023 in California is approximately -694 people, compared to the actual value in 2023 of 136,531. It seems very inaccurate and indeed, the average MSE for this model was 88,764,310,846.8214 with a MSE of 15,907,313,067 for this row. However, with data from 2007 – 2023 only, I am not sure I have enough information to smooth predictions for trends in the data that create the kind of exponential curves shown above in the non-linear graphs of overall homelessness in California or Los Angeles. In the ARIMA2 model, technically a Seasonal Autoregressive Integrated Moving Average + exogenous variables (SARIMAX) model, the predicted overall homelessness for California is 78,561. This value is much closer to our actual value. It seems like the ARIMA2 model might be the best predictor for our data after all. Given new data from years of collecting the model can become more precise as to how the 'seasonality' of years effects the curvature of the data to create this trend, which is observed as an exponential curve in the data.

Bilbliography

- Goines, David Lance. “Anatole France.” Poster accessed at [https://www.goines.net/Gallery/gal_xtra/007_anatole_france.gif] in September 2024.
- Ludden, Jennifer. 06/28/2024. “The Supreme Court says cities can punish people for sleeping in public places.” *Morning Edition.* National Public Radio Broadcast, accessed at [https://www.npr.org/2024/06/28/nx-s1-4992010/supreme-court-homeless-punish-sleeping-encampments] in September 2024.
- Wikipedia article: [https://en.wikipedia.org/wiki/Affordable_Care_Act], accessed in 2024.

Databases

- Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group. “Table 1: Total All Payers State Estimates by State of Residence (1991 - 2020) - Personal Health Care (Millions of Dollars).” [https://www.chcf.org/publication/2023-edition-california-health-care-spending/#related-links-and-downloads]
- California Government dataset, https://www.dir.ca.gov/OPRL/CPI/CPICalculator/CpiCalculator.aspx, accessed 2024.
- Department of Housing and Urban Development [https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/], accessed 2024.
- U.S. Bureau of Labor Statistics, Unemployment Rate in California [CAUR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CAUR, June 2, 2024.
- U.S. Federal Housing Finance Agency, All-Transactions House Price Index for California [CASTHPI]. Retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CASTHPI, June 2, 2024.
- The World Bank, World Development Indicators (2024). *US Health Indicators; Poverty and health indicators*. Retrieved from [https://databank.worldbank.org/source/world-development-indicators] in 2024.