Lars Finlayson 10/22/2024

Question 1: Preparing A Dataset in the Real World-

While browsing the Bureau of Labor Statistics site, I decided that the dataset that I wanted to work with was the one titled “Consumer Price Index for All Urban Consumers (CPI-U)” with the subtitle “All items in U.S. city average, all urban consumers, not seasonally adjusted," has I.D. listed as CUUR0000SA0, and I chose to include the years from 1913 to 2024 which is the furthest back the site collects data to the most recent. The dataset mainly deals with monthly Consumer Price Index values from January to December of each year as the labels of the columns and the years these values were collected. The Consumer Price Index measures inflation by tracking changes in the price of goods and services. While working with the data initially, I created an additional column called “CPI,” which I used to represent the average Consumer Price Index across all months of the year. This acted as the target variable for the predictive models. That was just the beginning of preparing the data. First I loaded the data into a pandas data frame. Then I dropped the missing values using dropna() to avoid errors during the modeling process, and all values were converted to numeric values. The target variable “CPI” was created by averaging the monthly CPI values across the 12 months which provided a single CPI value per year and allowed the model to focus on predicting the overall trend in consumer prices rather than just monthly fluctuations. This averaging helps to reduce noise and capture the general inflation trend. I scaled the data to ensure all features contribute equally to the model and applied PCA to reduce the dimensionality of the feature set, which simplifies and helps prevent overfitting. Lastly I split the dataset into training and testing sets using an 80-20 split to evaluate the model on unseen data.

Question 2: Define your problem-

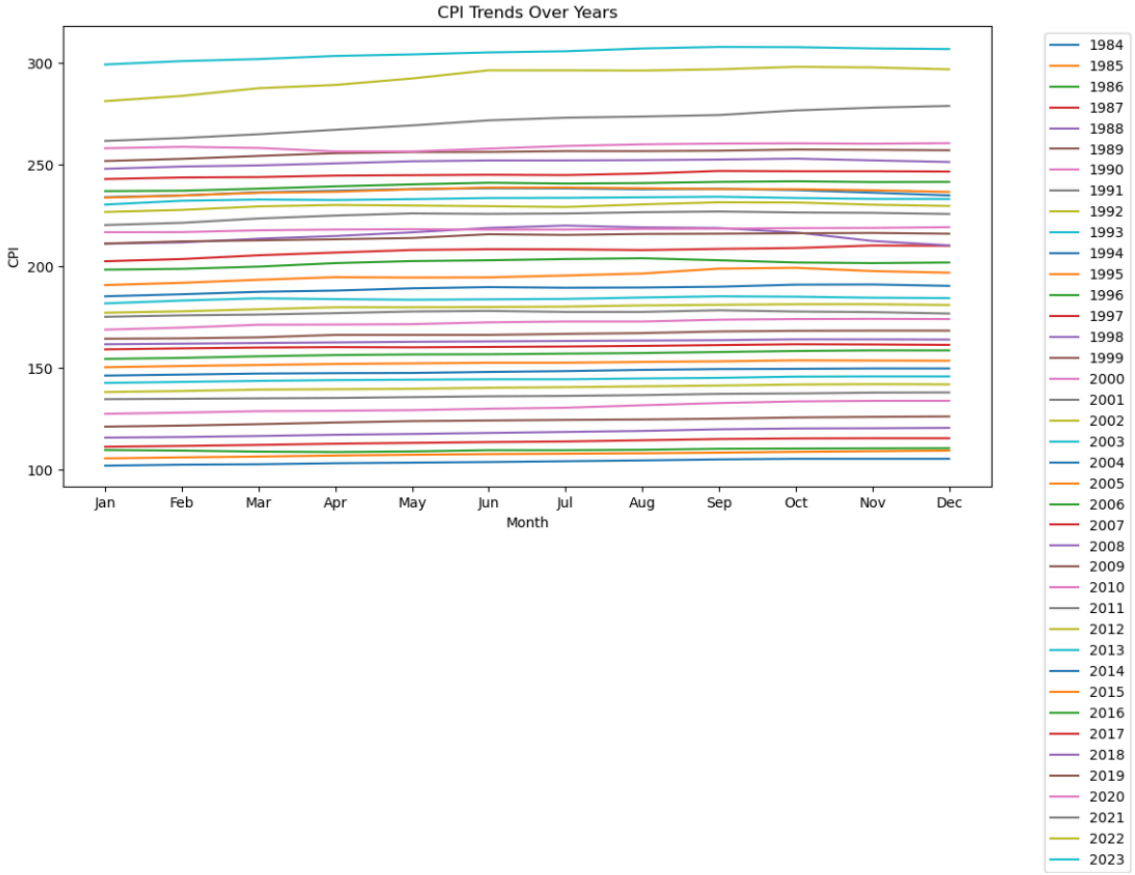
The problem that I decided to investigate during this analysis is to predict the annual average Price Index based on monthly CPI data from January to December. This is framed as a regression problem because CPI is a continuous variable. This predictive analysis can provide insight into inflation trends influencing economic policy decisions and consumer behavior. This problem definition was based on a few factors while exploring the data. Data analysis revealed that CPI fluctuates and shows seasonal trends in monthly CPI values. Some months may exhibit higher CPI due to holidays while others may show a lower CPI. Due to these patterns being shown using monthly CPI data to predict the annual average would be a valid approach. Also, a correlation analysis of monthly CPI values showed a strong result between adjacent months, confirming that price trends persist throughout the year. This makes using monthly CPI data as a feature for predicting the annual average CPI a valid approach. I used a regression model based on monthly data to predict the average annual CPI. I tested several regression models to determine their effectiveness, including linear, ridge, Lasso, random forest, KNN, and polynomial regression. Each of these models offers a different advantage in different situations. For example, a linear model can easily interpolate and provide a baseline, while polynomial regressions can find more complex relations. I have added some graphs and tables that helped me decide on a Regression problem.

Summary Statistics: The deceptive statistics of monthly CPI values show variations in consumer prices for different months. The standard deviation highlights which months tend to have more volatility in prices.

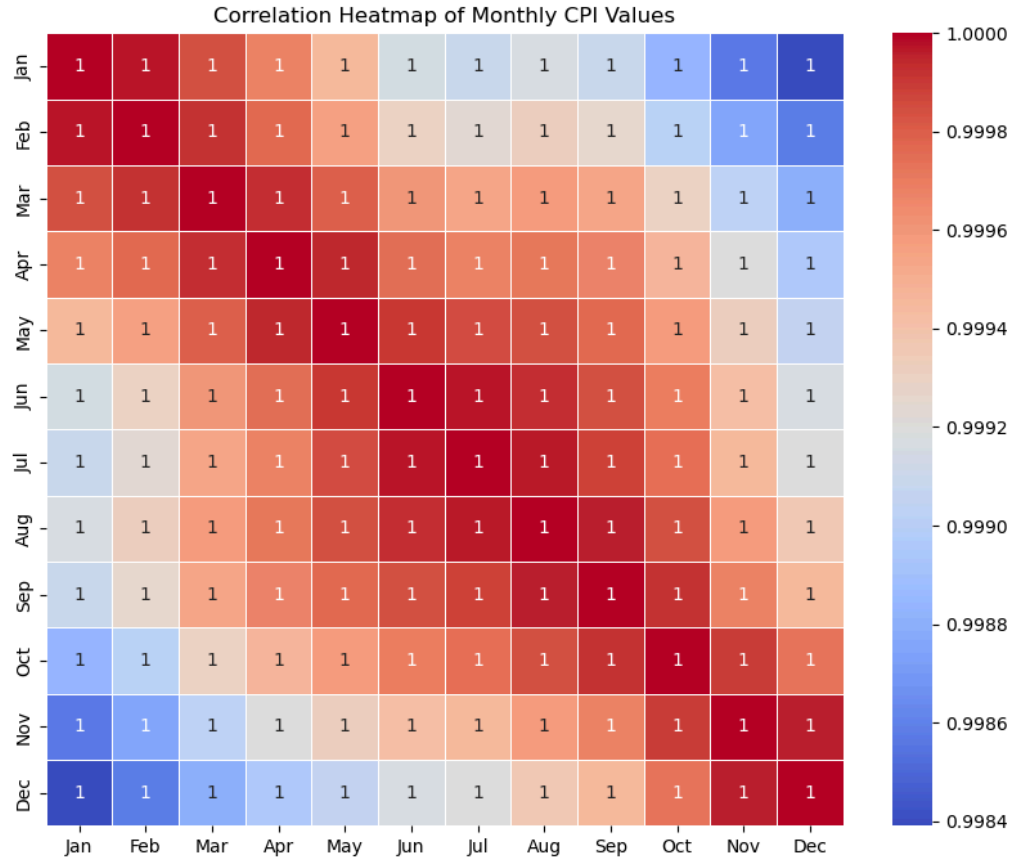
	Jan	Feb	Mar	Apr	May	Jun
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000
mean	187.145550	187.914900	188.824625	189.481125	190.057450	190.655700
std	52.921196	53.270775	53.665356	53.898105	54.223108	54.563113
min	101.900000	102.400000	102.600000	103.100000	103.400000	103.700000
25%	145.300000	145.800000	146.300000	146.550000	146.675000	147.100000
50%	183.450000	184.650000	185.800000	185.900000	186.300000	186.700000
75%	231.136750	232.805000	233.609500	233.548000	234.160000	234.713750
max	299.170000	300.840000	301.836000	303.363000	304.127000	305.109000

	Jul	Aug	Sep	Oct	Nov	Dec
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000
mean	190.892825	191.246800	191.730850	191.930875	191.743725	191.515825
std	54.584329	54.569163	54.545126	54.506762	54.301565	54.148496
min	104.100000	104.500000	105.000000	105.300000	105.300000	105.300000
25%	147.400000	147.950000	148.325000	148.550000	148.725000	148.725000
50%	186.650000	187.050000	187.550000	187.950000	187.750000	187.300000
75%	234.759500	234.870750	235.098000	234.517750	233.839500	233.489750
max	305.691000	307.026000	307.789000	307.671000	307.051000	306.746000

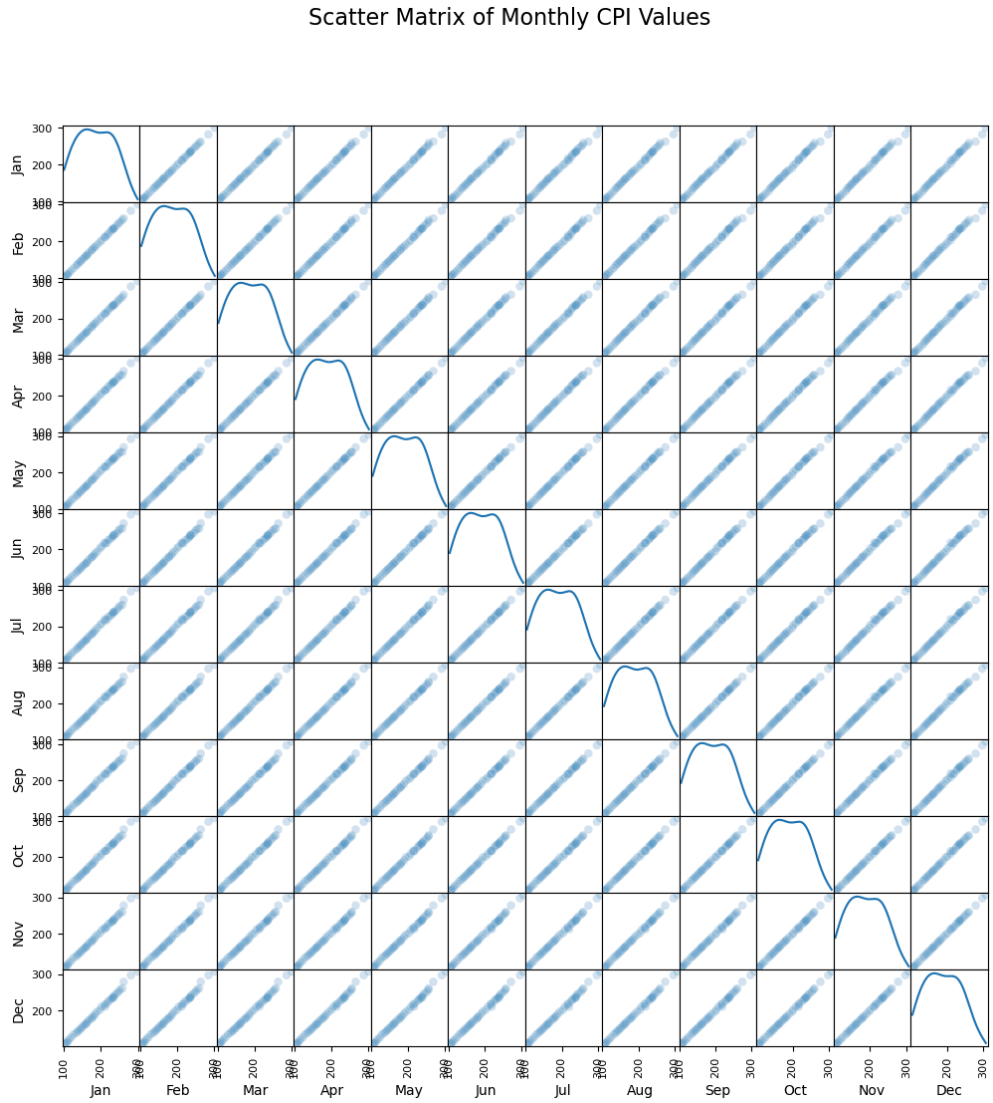
Times Series Plot: This time series plot of monthly CPI values can reveal consistent seasonal patterns in infection. For example, December shows higher CPI values, probably because of increased consumer spending during the holiday season.



Correlation Heatmap: The heatmap shows a strong positive correlation between monthly CPI values, showing that inflation persists across consecutive months. This supports my use of monthly CPI to predict the overall annual CPI.



Scatter Matrix: The scatter matrix shows that the monthly CPI values tend to have linear relationships, confirming the heatmap observations. These relationships help to predict the overall annual monthly CPI using monthly data.



A classification could be performed here by categorizing years as high or low inflation based on CPI thresholds. This could be done by recognizing that months such as December or January have a more significant impact on predicting the annual CPI than others. Redefining this problem could focus on these important months, simplifying the model and improving accuracy. On the other hand, the problem could be reframed as a classification task that could predict whether a year will experience high or low based on CPI thresholds.

Question 3: Justifying Model Selections- Using such a large number of linear and nonlinear models to test with and regularization was done to determine the best model for predicting CPI based on historical data. I tested multiple models to ensure that I was thorough enough and that I was positive in my selection based on model performance, which showed an effective approach to finding relationships between monthly CPI values and overall trends in inflation.

The predictors used in the analysis were the monthly CPI values from January to December. These features can capture trends throughout the year. CPI tends to show a high correlation across months so Principal Component Analysis, an unsupervised learning technique, was applied to reduce the dimeriality and address multicollinearity. The feature space was reduced from 12 months to 5 components and ensured the model focused on the most significant patterns in the data. I used linear regression as a baseline model to evaluate the simple linear relationship between monthly CPI and average CPI. Ridge Regression is a regularized version of linear regression, which helps address multicollinearity by penalizing large coefficients. Lasso Regression is a regularized model that also performs feature selection by shrinking some coefficients to zero, simplifying the model. Random Forest is a non-linear ensemble model that captures complex interactions between monthly CPI values. Polynomial Regression was used to capture non-linear relationships through higher-order terms. Gradient Boosting is a powerful boosting model that sequentially improves its predictions. KNN regression is a simple algorithm that predicts values based on the proximity of observations. I evaluated the models using cross-validation, mean-squared error, and R^2 to assess the models' performance on the data. I used K-Fold-Cross-Validation to assess the performance of the models, which gave me Cross-validation, Test MSE, and Test R^2. The Mean Squared Error measures the average difference between the predicted and actual CPI values. A lower MSE can be an indicator of better model performance. R^2 measures the proportion of variance in the CPI explained by the model. A higher R^2 indicates that the model fits the data well. Here are the table and values of the model's performances.

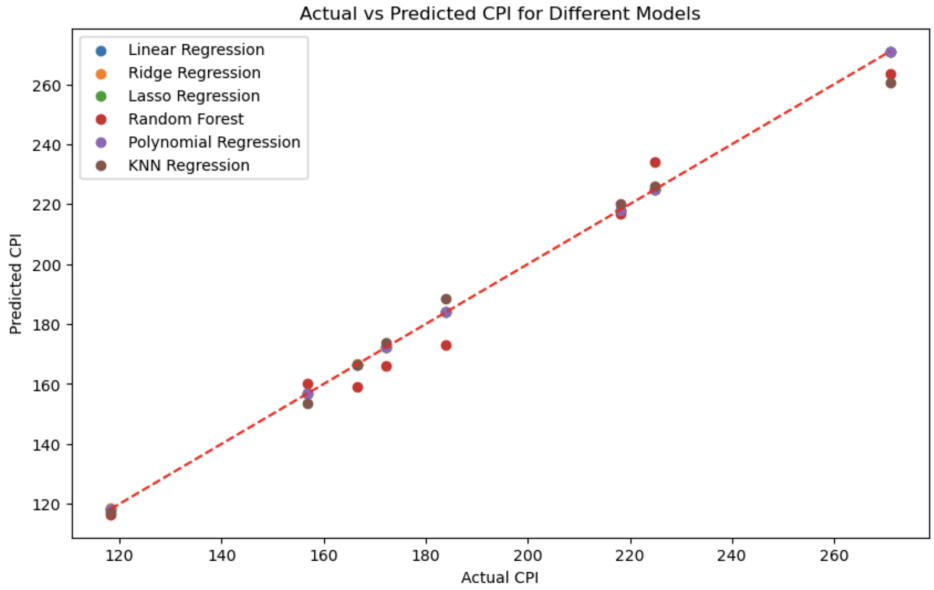
	Linear Regression	Ridge Regression	Lasso Regression	\
Cross-Validation R²	1.000000e+00	0.999989	1.000000	
Test MSE	1.966057e-08	0.012621	0.000776	
Test R²	1.000000e+00	0.999994	1.000000	

	Random Forest	Polynomial Regression	KNN Regression
Cross-Validation R²	0.953242	1.000000e+00	0.955653
Test MSE	46.972676	1.966057e-08	18.178325
Test R²	0.976070	1.000000e+00	0.990739

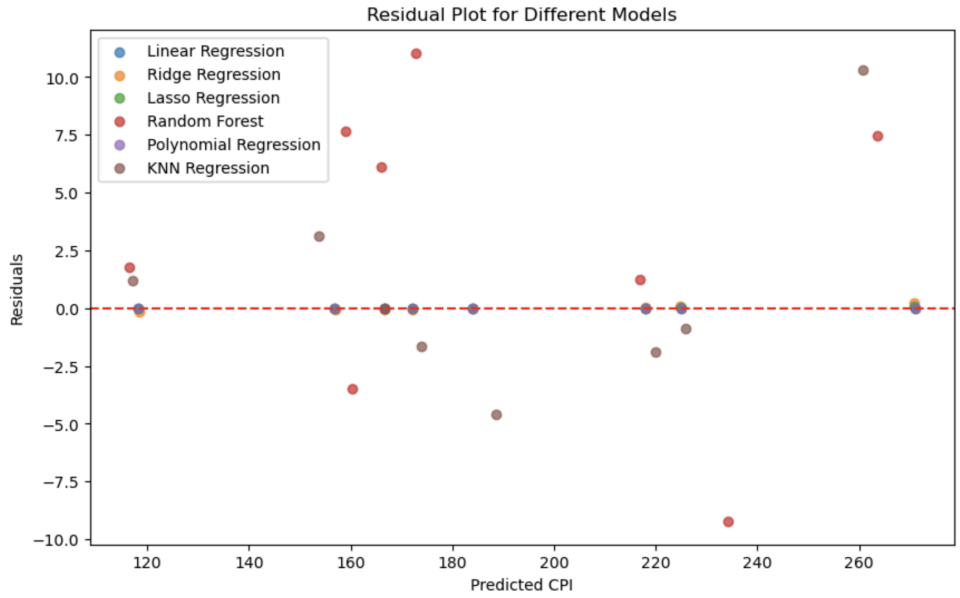
These results show that linear, polynomial, and lasso regression performed very well, with little test MSE and R2 being close to 1. The only other model that performed well was Ridge, but Ridge had a higher error. Polynomial regression performs well but is more complex than the other models. KNN and Random Forest did not perform well as shown by a high error. Based on these results, I used Lasso regression despite its slightly higher errors. I decided to do this because while the error is not as low as linear regression, it is still very low. On top of that lasso regression in general is more robust against overfitting and lasso can perform feature reflection by shrinking some coefficients to zero which could simplify the model and make it more interpretable. Both lasso and linear regression had perfect R^2 values, so they are equal.

Question 4: Drawing Data-Given Conclusions-
Some visualizations of the predictive analytics results:

Actual vs Predicted CPI Plot: This plot showcases how well the different models predict CPI values by comparing the actual CPI against the model predictions. Linear, Polynomial, and Ridge are hard to see because they performed well with Lasso, and the points are on each other.



Residual Plot: This plot shows how the differences between actual and predicted values are distributed and which models exhibit the least error. While the Lasso and Linear points are still completely overlapped, there is some separation now with the ridge regression.



As shown in the two graphs above and the explanations I gave in the previous section, Lasso is the best model for this dataset. Although there are not many graphs, I believe that only these two are needed to show that Linear, Lasso and, Polynomial are the best models and because of the explanations that I have already given these two graphs are all that are needed to show the lasso regression model is the best. In the long-term use of the model, the performance of the Lasso regression model using Mean Squared Erro and R^2 as more data is added to the dataset. Periodically, as more data comes in, retraining the model could be a good idea to maintain its accuracy and relevance as economic conditions change worldwide. Because I chose Lasso regression, this could mean that because of its feature selection it may be able to adapt better to changing relationships between monthly CPI values over time making the model more robust.