LINEAR REGRESSION AND TIME SERIES FORECASTING

Bharath Kumar Karre, Sumanth Pobala, Swathi Priya Soogoor

***Abstract*— Global warming is one the major environmental issues in the present world. The average global temperature is increasing annually and there are many reasons for this increase. The report observes different toxic emissions in the atmosphere and analyze them to find out which have significant contribution to the increase in temperature. The other analysis conducted in this report is time series implementation on Avocado prices and the future forecast for the prices.**

***Key Terms*— Global Warming, Linear Regression, Variance Inflation Factor, Multi-collinearity, Stats models, Ordinary Least Squares, Test-Train split, Cross Validation, Time Series Analysis, Rolling Window, Seasonal Decomposition, ARIMA Model, SARIMA Model, Future Forecasting.**

I. INTRODUCTION

The drastic increase of the average global temperatures in the present day sets a question that what is the major cause of this increase? We have taken into consideration, the toxic emissions in the atmosphere and perform a linear regression analysis to conclude the significant factors contributing to increase in temperature.

The other analysis is time-series forecasting on Avocado prices. Avocado is one of the most-liked fruits of the present generation. The price trends of this most-liked fruit in various regions are analyzed with the data of previous three years. Also, the price forecast for the future is also observed.

II. HYPOTHESIS

A. Global Warming

The hypothesis for the analysis of factors for the increase in temperature is based on the coefficients of the factors in the linear regression. It is as follows:

Null Hypothesis: There is no significant effect of the emissions on the global temperature. (In other words, the coefficients of the emissions in the regression model are zero)

Alternate Hypothesis: There is at least one emission which has significant effect on the global temperature.(The coefficients of at least one emission is not zero)

Based on the hypothesis and the p-value, we get from the regression model, we either fail to reject the null hypothesis or we reject the null hypothesis.

B. Avocado Prices

In the time series forecasting of the Avocado prices, we check for the trends of the prices over a certain period of time. While conducting the analysis, depending on our data set, we would like to find answers for the following questions:

a) Which region has the highest and lowest prices for Avocados in the United States?

b) What is the price comparison between organic and traditional Avocados?

c*)* How is the variation of prices and what is future trend of prices?

As the analysis is moving forward, we aim to find appropriate conclusions for our hypothesis.

III. DATA SETS AND METHODS

A. Global Warming

The global warming data set consists of different emissions namely Methyl Iodide(MEI), Carbon di-oxide(CO2), Methane(CH4), Nitrous Oxide(N2O), Trichlorofluoromethane (CFC 11), Dichlorodifluoromethane(CFC 12), Total Solar Irradiance(TSI), Aerosols. The amount of emissions in the atmosphere is given for every month starting from May,1983 to December 2008. The corresponding global temperature is also given for every month from 1983-2008.

The data is loaded into a data frame in jupyter notebook and the data is checked for the shape, size, data types and null values. We got to know that there are no null values in the data set.

a) Correlation

To find the correlation among the factors, a heat map is used. A correlation table and heat map are generated. The heat map is as follows:

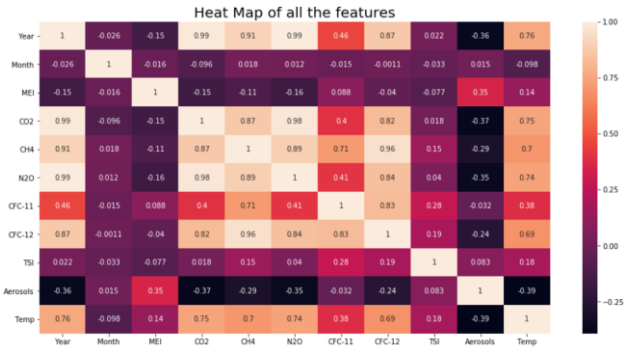

Fig. 1. Heat Map for correlation of the factors

From the heat map, there is a high correlation of 0.75 for CO2 followed by N2O with 0.74. The least correlation being -0.39 for Aerosols.

b) Multicollinearity

The interaction between the factors is checked with the variation inflation factor(VIF). This is generally termed as multicollinearity. The optimal VIF should be almost equal or less than one. In the analysis, for model 1, we considered all the factors for which the VIF for CFC 12 is high indicating high multicollinearity with other factors. So, CFC 12 is eliminated from the factors and model 2 is considered. For model 2, VIF for N2O is high. Repeating the process of elimination, for model 3, we observed high VIF for CH4, after eliminating N2O. Finally, in model 4, after eliminating CH4, all the VIF are near to one.

The remaining factors which does not have multicollinearity, or the multicollinearity is least are MEI,CO2,CFC 11,TSI and Aerosols.

c) StatsModels

Statsmodels is a Python package that allows users to explore data, estimate statistical models, and perform statistical tests. An extensive list of descriptive statistics, statistical tests, plotting functions, and result statistics are available for different types of data and each estimator.

Using all the factors in data, excluding year and month, a regression model named as model 1, is built, and the summary is as follows:

OLS Regression Results

Dep. Variable:	Temp	R-squared:	0.744
Model:	OLS	Adj. R-squared:	0.737
Method:	Least Squares	F-statistic:	108.6
Date:	Wed, 04 Dec 2019	Prob (F-statistic):	8.21e-84
Time:	22:00:33	Log-Likelihood:	303.02
No. Observations:	308	AIC:	-588.0
Df Residuals:	299	BIC:	-554.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-127.6958	19.191	-6.654	0.000	-165.462	-89.929
MEI	0.0663	0.006	10.722	0.000	0.054	0.078
CO2	0.0052	0.002	2.375	0.018	0.001	0.010
CH4	6.371e-05	0.000	0.128	0.898	-0.001	0.001
N2O	-0.0169	0.008	-2.161	0.032	-0.032	-0.002
CFC1	-0.0073	0.001	-4.980	0.000	-0.010	-0.004
CFC2	0.0043	0.001	4.875	0.000	0.003	0.006
TSI	0.0959	0.014	6.844	0.000	0.068	0.123
Aerosols	-1.5818	0.210	-7.535	0.000	-1.995	-1.169

Omnibus:	6.703	Durbin-Watson:	0.978
Prob(Omnibus):	0.035	Jarque-Bera (JB):	8.299
Skew:	0.191	Prob(JB):	0.0158
Kurtosis:	3.708	Cond. No.	8.58e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.58e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Fig. 2. OLS Summary Table for Model 1

To check for significant factors, we need to take a look at the p-value. Here, in model 1, the p-value for Methane(CH4) is 0.898 which is greater than 0.05 and implies that the factor CH4 is not significant.

The mean square error(MSE), root mean square error(RMSE), R square, R square adjusted values, AIC, BIC are as follows:

```
MSE:  0.008184260128792465
RMSE:  0.09046690073608395
R2:  0.7439939571287729
R2adj:  0.7371442971188404
AIC:  -588.0409430107441
BIC:  -554.4700449639819
```

A regression plot is observed for the predicted values of the temperature and the actual values. The plot is as follows:

Fig. 3. Regression Plot for StatsModels Model 1

The factor CH4 is eliminated from the analysis and another built is using the remaining factors. This model is named as model 2. The summary table is as follows:

OLS Regression Results

Dep. Variable:	Temp	R-squared:	0.744
Model:	OLS	Adj. R-squared:	0.738
Method:	Least Squares	F-statistic:	124.5
Date:	Wed, 04 Dec 2019	Prob (F-statistic):	7.14e-85
Time:	22:08:48	Log-Likelihood:	303.01
No. Observations:	308	AIC:	-590.0
Df Residuals:	300	BIC:	-560.2
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-127.6250	19.151	-6.664	0.000	-165.313	-89.937
MEI	0.0662	0.006	10.783	0.000	0.054	0.078
CO2	0.0052	0.002	2.380	0.018	0.001	0.009
N2O	-0.0166	0.007	-2.227	0.027	-0.031	-0.002
CFC1	-0.0073	0.001	-4.998	0.000	-0.010	-0.004
CFC2	0.0043	0.001	5.055	0.000	0.003	0.006
TSI	0.0958	0.014	6.854	0.000	0.068	0.123
Aerosols	-1.5830	0.209	-7.559	0.000	-1.995	-1.171

Omnibus:	6.603	Durbin-Watson:	0.976
Prob(Omnibus):	0.037	Jarque-Bera (JB):	8.077
Skew:	0.192	Prob(JB):	0.0176
Kurtosis:	3.694	Cond. No.	5.69e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.69e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Fig. 4. OLS Summary Table for Model 2

All the p-values are less than 0.05 indicating the significance. The mean square error(MSE), root mean square error(RMSE), R square, R square adjusted values, AIC,BIC are as follows:

```
MSE:  0.00818470866133402
RMSE:  0.09046937968911924
R2:  0.7439799269001077
R2adj:  0.7380061251944435
AIC:  -590.0240637528502
BIC:  -560.1832654890617
```

The regression plot for model 2 is shown below:

Fig. 5. Regression Plot for StatsModels Model 2

The significant factors affecting the temperature are known, but we cannot conclude that the regression model obtained is the best fit. So, another model named model 3 is built using the factors which does not have multicollinearity.

Model 3 is built using the factors MEI,CO2,CFC 11,TSI and Aerosols.

OLS Regression Results

Dep. Variable:	Temp	R-squared:	0.719
Model:	OLS	Adj. R-squared:	0.714
Method:	Least Squares	F-statistic:	154.5
Date:	Wed, 04 Dec 2019	Prob (F-statistic):	5.08e-81
Time:	22:09:32	Log-Likelihood:	288.64
No. Observations:	308	AIC:	-565.3
Df Residuals:	302	BIC:	-542.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-138.8416	19.881	-6.984	0.000	-177.964	-99.719
MEI	0.0683	0.006	10.735	0.000	0.056	0.081
CO2	0.0099	0.001	19.151	0.000	0.009	0.011
CFC1	-2.964e-05	0.000	-0.094	0.925	-0.001	0.001
TSI	0.0992	0.015	6.815	0.000	0.071	0.128
Aerosols	-1.7201	0.216	-7.970	0.000	-2.145	-1.295

Omnibus:	11.534	Durbin-Watson:	0.912
Prob(Omnibus):	0.003	Jarque-Bera (JB):	15.097
Skew:	0.312	Prob(JB):	0.000527
Kurtosis:	3.887	Cond. No.	5.23e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.23e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Fig. 6. OLS Summary Table for Model 3

If the p-values here are given a keen observation, the value for CFC 11 is 0.925 which greater than 0.05 indicating the insignificance. Due to this, we have eliminated the CFC-1 feature and performed regression on the remaining features. These features turned out to be Model 4.

```
MSE:  0.008985275872117466
RMSE:  0.09479069507139119
R2:  0.7189379511491043
R2adj:  0.7142846059694536
AIC:  -565.2816246669481
BIC:  -542.9010259691066
```

The regression plot for model 3 is as follows:

Fig. 7. Regression Plot for StatsModels Model 3

As model 3 has one insignificant factor, we are building another model, model 4 eliminating the CFC 11 factor. The summary table is:

OLS Regression Results

Dep. Variable:	Temp	R-squared:	0.719
Model:	OLS	Adj. R-squared:	0.715
Method:	Least Squares	F-statistic:	193.8
Date:	Wed, 04 Dec 2019	Prob (F-statistic):	3.44e-82
Time:	22:10:09	Log-Likelihood:	288.64
No. Observations:	308	AIC:	-567.3
Df Residuals:	303	BIC:	-548.6
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-138.2575	18.849	-7.335	0.000	-175.349	-101.166
MEI	0.0682	0.006	10.897	0.000	0.056	0.080
CO2	0.0099	0.000	21.205	0.000	0.009	0.011
TSI	0.0988	0.014	7.156	0.000	0.072	0.126
Aerosols	-1.7212	0.215	-7.998	0.000	-2.145	-1.298

Omnibus:	11.490	Durbin-Watson:	0.911
Prob(Omnibus):	0.003	Jarque-Bera (JB):	14.999
Skew:	0.312	Prob(JB):	0.000553
Kurtosis:	3.883	Cond. No.	4.89e+06

Fig. 8. OLS Summary Table for Model 4

All the p-values are less than 0.05 implying all the factors considered in the above model are significant. The mean square error(MSE), root mean square error(RMSE), R square, R square adjusted values, AIC,BIC are as follows:

```
MSE:  0.008985537581110689
RMSE:  0.09479207551852997
R2:  0.7189297648154975
R2adj:  0.7152192666612467
AIC:  -567.2726538576824
BIC:  -548.6221549428145
```

The regression plot for the model 4 is:

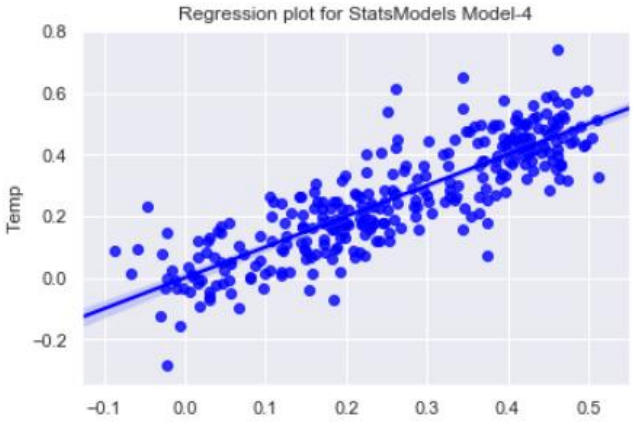

Fig. 9. Regression Plot for StatsModels Model 4

d) Correlation between Actual and Predicted Values of the Temperature

To show the correlation between the predicted and actual values of the temperature, we have plotted a joint plot as below:

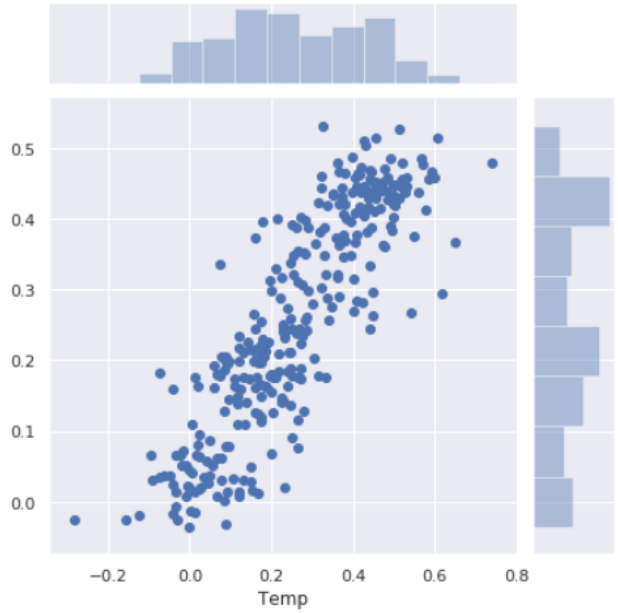

Fig. 10. Joint Plot

From the above plot, we can infer that there is a positive correlation between the actual and predicted values. The correlation coefficient is 0.86255.

e) SKLearn Linear Regression with Test- Train Split

As part of test-train split linear regression, we implemented Sklearn test-train split linear regression. The test size is 20% of the original data. The random state is 5.

By using test and train data set, we have built all the four models considered in stats models with same factors for the respective models.

The mean square error(MSE) values and the regression plot for the four models built using test-train split:

Model 1

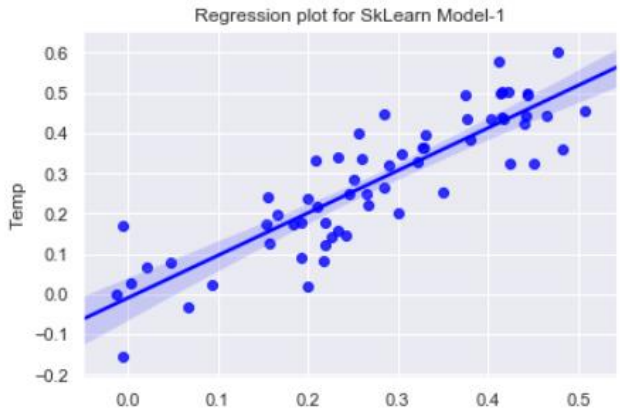

Fig. 11. Regression Plot for SKLearn Model 1

MSE is 0.006844

Model 2

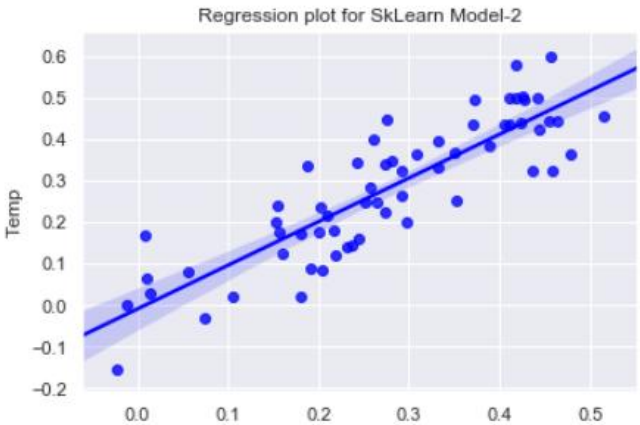

Fig. 12. Regression Plot for SKLearn Model 2

MSE is 0.006910

Model 3

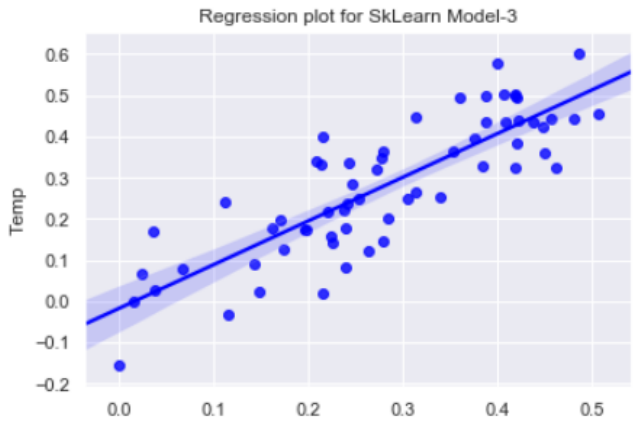

Fig. 13. Regression Plot for SKLearn Model 3

MSE is 0.008092

Model 4

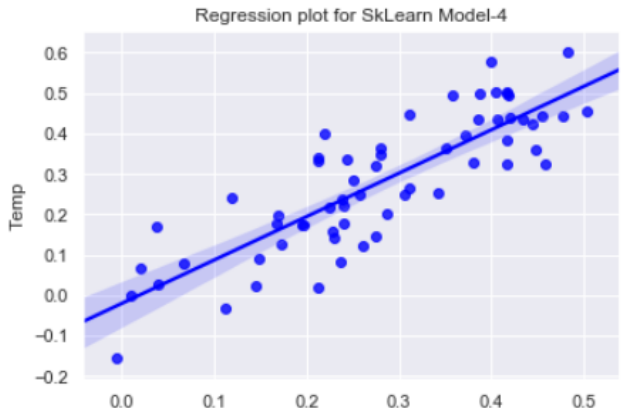

Fig. 14. Regression Plot for SKLearn Model 4

MSE is 0.008092

f) Linear Regression using Cross Validation

For the cross validation, we have split the data into 10 folds. In every fold, we have selected a fold as Test data and remaining as train data. In each experiment, the test fold varies. In this way, we have performed the cross validation and attained an average RMSE value.

The models are built using cross validation scores. The factors for each model are the same which are mentioned above. The cross-validation scores for each of the model are mentioned below.

Model 1 : 0.010077
Model 2 : 0.009984
Model 3 : 0.010951
Model 4 : 0.010077

The best model is observed using the MSE and RMSE values which would be discussed in the results.

B. Avocado Price Prediction

Avocado prices are predicted based on the following data. Some of the relevant factors are described below:

Date - The date of the observation
Average Price - The average price of a single avocado
Type - Conventional or Organic
Year - The year
Region - The city or region of the observation
Total Volume - Total number of avocados sold
4046 - Total number of avocados with PLU 4046 sold
4225 - Total number of avocados with PLU 4225 sold
4770 - Total number of avocados with PLU 4770 sold

Other predictors in data set that are less significant in the analysis are Total Bags, Small Bags, Large Bags, XL Bags. These are not considered for the analysis.

The data set chosen is weekly data starting from last week of December 2015 to first week of January 2018.

The implementation of ARIMA model, SARIMA model is done for the prediction of trend of the prices and also to forecast the prices into the future.

The data is loaded into a data frame in jupyter notebook and the data is checked for the shape, size, data types and null values. Generally, python can detect a particular date and time format which is yyyy-mm-dd. The date format in the chosen data set is already in the python readable format. So, by using the sort_index function, we set the index to date and continue the further analysis.

a) Rolling Window

Rolling window function is used to check particular number of data points from a given data point. Here, we check the rolling window of the prices of Avocados 150 days in to the past to estimate the trend of the prices. The rolling window for Avocado prices is as shown below.

Fig. 15. Rolling window of prices

b) Data Exploration

As mentioned in the hypothesis, we want to explore the data and find out answers to research questions. One of such

explorations is to find out the price variations in conventional and organics Avocados. To achieve, this we plotted a line plot for the prices for conventional and organic Avocados in the whole US. The plot obtained is as below.

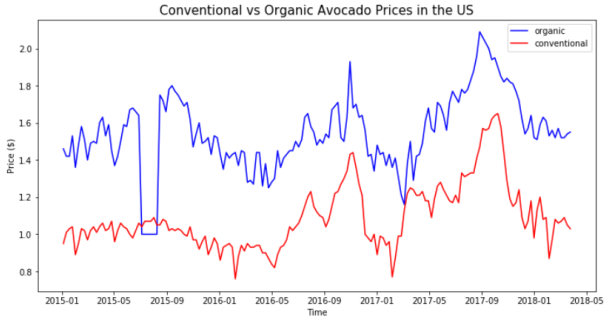

Fig. 16. Comparison of Prices for Organic and Conventional Avocados

We also want to find, how the prices for conventional Avocados in New York, Boston, DallasFortWorth and in total US vary. The results obtained are as follows:

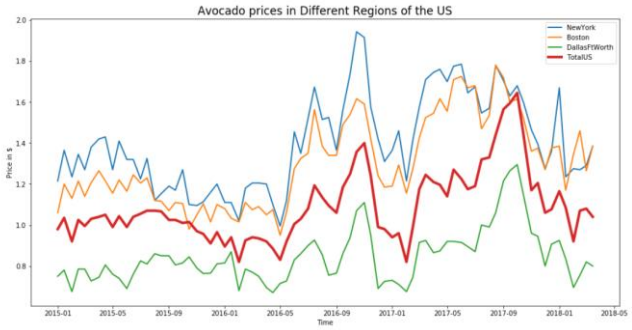

Fig. 17. Prices for Conventional Avocados in different regions

c) Time Series Analysis

To implement the time series analysis, first we take a look at the variation of existing prices. The monthly variation of the average prices of Avocados is shown below:

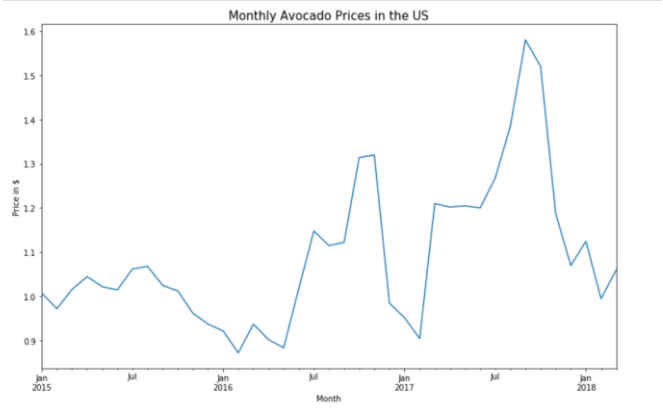

Fig. 18. Monthly Variation of prices for Avocados in US

The next step in implementation of time series is making the data stationary. To achieve this, seasonal decomposition function is used. The data is disaggregated into three components- trend component, seasonal component and residual component.

Trend Component: It generally gives the slow-moving overall level of the series.

Seasonal Component: It captures the patterns which are repeated every season.

Residual Component: It is the left-over data other than trend and seasonal component.

The data after the seasonal decomposition is shown below.

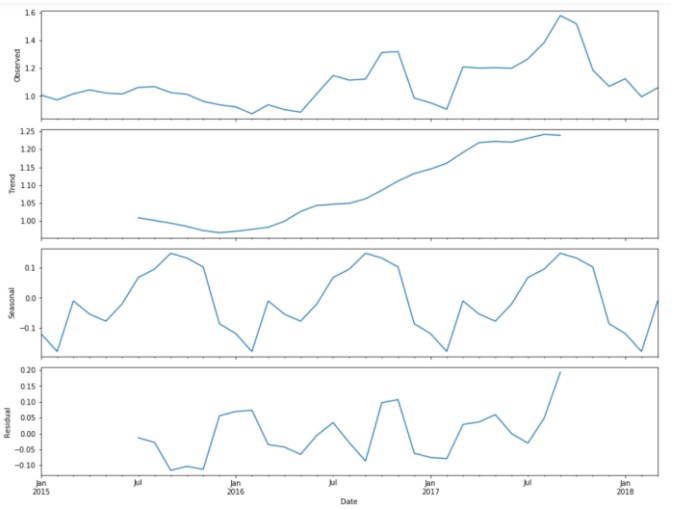

Fig. 19. Seasonal Decomposition of data

ARIMA Model

ARIMA is **A**uto **R**egressive **I**ntegrated **M**oving **A**verage. It uses a number of lagged observations of time series to forecast observations. The predictions using ARIMA Model is as shown below.

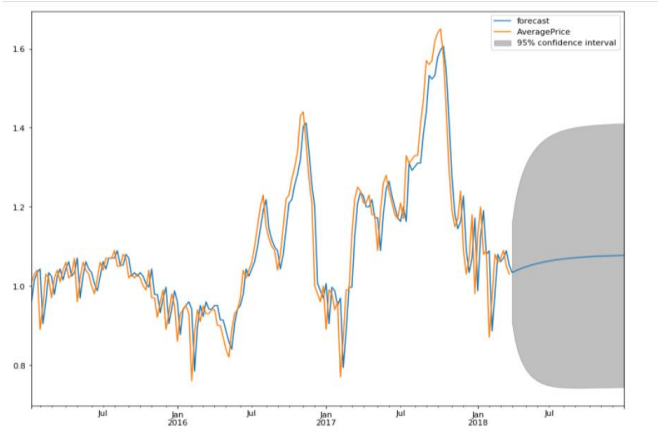

Fig. 20. ARIMA Model

SARIMA Model

SARIMA model is used to analyze and forecast data which have an additional seasonal component. The forecast using SARIMA model is shown below.

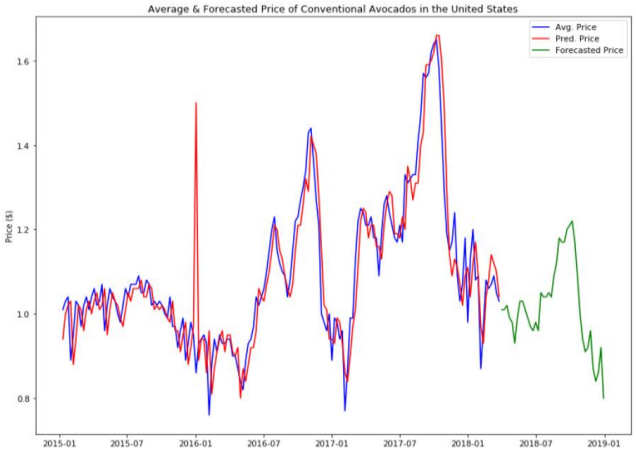

Fig. 21. SARIMA Model

MSE is 0.006040
RMSE is 0.077720

The results obtained from the SARIMA Model are as follows.

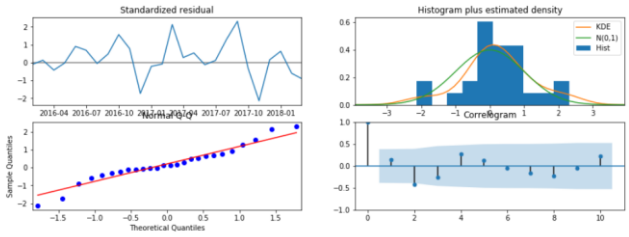

Fig. 22. Results from SARIMA Model

The variation of the true prices and predicted prices.

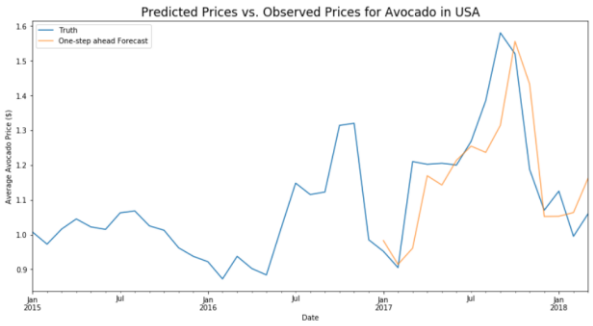

Fig. 23. Variation of actual and predicted prices for Avocados

Then we predicted the future prices of the Avocados using the existing prices.

IV. Results and Discussions

A. Global Warming

a) Correlation

The correlation for all the factors with the temperature is given below.

MEI	0.14
CO2	0.75
CH4	0.7
N2O	0.74
CFC-11	0.38
CFC-12	0.69
TSI	0.18
Aerosols	-0.39

The correlation for CO2 with temperature is high and Aerosols is low.

b) Multicollinearity

The VIF for all the factors are given below:

	Iter 1	Iter 2	Iter 3	Iter 4
Temperature	13454981	13294463	13286575	13284513
MEI	1.225	1.208	1.207	1.192
CO2	27.996	27.828	7.312	1.425
CH4	19.129	18.069	12.479	
N2O	61.037	39.813		
CFC-11	31.829	4.613	3.768	1.369
CFC-12	93.498			
TSI	1.140	1.140	1.134	1.133
Aerosol	1.354	1.329	1.318	1.316

c) Stats Models

We have built four models using different factors. The best model is based on the least RMSE value. The RMSE values of four models are given below.

Model 1	0.090466
Model 2	0.090469
Model 3	0.094790
Model 4	0.094792

Based on above values, Model 2 has least RMSE. Model 2 is better among all the models.

d) Sk-Learn train-test split

The model built can be checked with the RMSE values. As above mentioned, the model with least RMSE value is better amongst another models. The RMSE values of the models obtained with sk-learn train-test split is,

Model 1	0.006944
Model 2	0.006910
Model 3	0.008092
Model 4	0.008092

Model 2 is performing better among other models.

e) Linear Regression with cross validation
The RMSE values for models are:

Model 1	0.010077
Model 2	0.009984
Model 3	0.010951
Model 4	0.010077

Model 2 has the least RMSE and it is performing better among all the models.

B. Avocado Prices

a) Rolling Window
The rolling window of prices of Avocados shows that there is a fluctuation of prices.

b)Data Exploration
The prices of organic Avocados are expensive than the conventional Avocados. Also, there is a drastic fall of prices for organic Avocados in the period of July- August in 2015.

When the prices of conventional Avocados in the regions of New York, DallasFortWorth, Boston and Total US, New York has the higher prices and DallasFortWorth has lower prices than other regions comparatively.

c) Time Series Analysis
The monthly prices of Avocados are fluctuating, and the prices are rising in Fall of every year consistently and falling down at the end of each year.
The ARIMA model has accurately predicted the trend of the conventional prices of Avocados and when the future forecast is observed the prices are not increasing or decreasing, they are almost constant.

The SARIMA model is also appropriate except that it predicted a high rise in prices in January 2016, but the actual prices were low when compared to predicted prices. The future forecast using this model, tells that the prices are fluctuating, and the trend of the prices is similar to the past prices.

The histogram generated using SARIMA model depicts that the data is in 95% confidence interval and also, the data is normal. The quantile-quantile plot is almost a straight line, which is a good indication. The Autocorrelation plot shows that there is correlation between the data points.

The line plot between actual and predicted average prices for Avocados has deviated from the actual average prices but, however it has predicted the trend accurately.

The future forecast of the prices for three – four years has shown that the implementation of time series has been accurate. There is an increase of the prices every year and the trend of the prices remains to be the same.

V. Conclusions

Comparing all the models we built using linear regression, we reject the null hypothesis. So, there is at least one factor that significantly affect the global temperature. In conclusion, the temperature is significantly affected by Methyl Iodide(MEI), Carbon di-oxide(CO_2), Nitrous Oxide(N_2O), Trichlorofluoromethane(CFC_{11}), Dichlorodifluoromethane (CFC_{12}), Total Solar Irradiance(TSI),Aerosols.

From the time series implementation on Avocado Prices, we observed that,

- The organic avocados are expensive than conventional avocados.
- There are some distinguishable patterns between organic and conventional avocados, but, however, there are most patterns that are similar between these two types.
- The prices in the year 2017 has been high for Avocados. The reason for this is that it may the economy or there might be any other reason.
- The prices of avocados are consistently high in fall. We can suggest that buying of Avocados before fall is an optimal step.
- Based on SARIMA model, we see a downward trend in the prices in the long run, however there is increase in trend in short run, but the prices are decreasing in the coming years.

VI. References

- http://www.hassavocadoboard.com/retail/volume-and-price-data
- https://data.world/data-society/global-climate-change-data
- https://data.world/datasets/global-warming
- https://www.e-education.psu.edu/meteo469/node/215
- https://skepticalscience.com/anthropogenic-global-warming-rate-Is-it-steady-for-last-100-years.html
- https://www.sciencedirect.com/science/article/pii/S0012825218303726

VII. Team Member Contribution Statement

We, members of Team Ice Hockey , hereby declare that the following work is done by the members specified.
Bharath Kumar Karre – Worked on Linear Regression with Statsmodels, Sklearn Test Train Split.
Sumanth Pobala – Sklearn, Cross-validation evaluation, Exploratory data analysis for Time Series data Interpretation.
Swathi Priya Soogoor – Implementing ARIMA, SARIMA models on Avocado Data for Time Series Forecasting.