TCS Wastewater Team

TATA CONSULTING SERVICES X CARNEIGIE MELLON UNIVERSITY

Yi Chen Huang, Zhaoyu Qiao, Wen Ling Chang, Hanyun Zhang, Haijing Fang
Heinz College, Carnegie Mellon University, Pittsburgh, PA, USA

Table of Contents

1. Executive Summary

[Project 1] Waste and Water Utility

Water and wastewater systems worldwide maintain and operate the sewer and storm water system network to ensure hygienic, healthy living conditions for residents. This network of pipelines requires daily inspection and maintenance. Utilities spend millions of dollars annually for this purpose and typically hire contractors to go through inspection processes. One of the inspection methods performed by contractors is sending a CCTV camera inside the pipeline and capture the video footage and then manually go through them to identify issues like cracks, blockages, damages etc. There is an opportunity to create AI/Computer Vision based system to automate the process of inspecting these CCTV camera feed. Our team reviewed a variety of related technical documentations and research papers, examined this market from various viewpoints and produced two market analysis reports—Stakeholder analysis report and competitive analysis report.

The stakeholder analysis report identifies all the parties involved in the sewage pipeline industry, which provide analysis that discusses the benefits and costs associated with these parties. The competitor analysis report takes deep dive into current companies with CCTV related ML focused platforms.

[Project 2] Electricity Demand Prediction

The importance of energy demand management has been more vital in recent decades as the resources are getting less, emission is getting more and developments in applying renewable and clean energies has not been globally applied.[1] Therefore, for both governments and private companies, it is essential to understand the energy consumption patterns, and control its supply-demand relationship.

As U.S. Energy Information Administration (EIA) provides state-level monthly electricity consumption data on their website, our team use this dataset as our base source to build prediction models. To increase prediction accuracy, our team also merged with several external features like temperature, wind speed, and population to explain the

patterns of electricity consumption. Our team build linear regression and artificial neural network models to forecast national and state-level monthly consumption by commercial, industrial, and residential sector. The focus of this project is to find out hidden relationships between electricity demand among each sectors and different climate, demographic and economic variables.

Source:

https://link.springer.com/article/10.1007/s12667-016-0203-y#:~:text=Demand%20forecasting%20plays%20a%20vital,power%20production%20and%20distribution%20systems.

2. Workflow Summary

2.1 Project Organization

Project Managers:

Organization	Name
Heinz College	Professor Christopher Kowalsky
Tata Consultancy Services	Consultant Mehul Shah

Team members:

Name	Title & Responsibilities
Zhaoyu Qiao (Joy)	**Title:** Project Manager, Risk Manager
	Responsibilities: Point of contact to manager Mehul and external stakeholders, identify the risks of the project, Organize and format the final report
Yi Chen Huang (Maggie)	**Title:** Project Manager
	Responsibilities: Point of contact to manager Chris and external stakeholders, Control the progress of the overall project, Organize and format the final report
Hanyun Zhang (Hanyun)	**Title:** Chef System Architect Manager, Project Documentation Manager
	Responsibilities: Design the pipelines for our electricity demand prediction model, Organize all mid-term deliverables
Wen Ling Chang (Linda)	**Title:** Quality Assurance Manager, Project Documentation Manager
	Responsibilities: Design the pipelines for our electricity demand prediction model, Code quality management, Organize and format all mid-term deliverables
Haijing Fang (Haijing)	**Title:** Quality Assurance Manager
	Responsibilities: Design the pipelines for our electricity demand prediction model, Code quality management

2.2 Project Scheduling

Time & Duration	Plan & Schedule	Accomplishment	Details
Week 1: 6/8-6/12	1. Warm up and icebreaking of our team 2. Make plans for Week 2 and Week 3 3. Understanding the project scope and the CCTV Inspection on Sewage Industry backgrounds	1. Plan for the next two Weeks 2. Google Drive setting 3. Team rules setting 4. Industry research and discussion	1. Setting up meeting time with TCS manager, Mehul, every week 2. Setting up meeting time with Professor every week 3. Setting up internal meeting time every day 4. Having learning lecture with TCS industry specialist, Avishek, in order to solve our confusion 5. Plan to complete Project 1: Stage 1-Stakeholder Analysis Report on Week 2 and Stage 2-Competitor Analysis Report on Week 3
Week 2: 6/15-6/19	1. Research the CCTV inspection industry thoroughly 2. Complete Project 1 Stage 1-Stakeholder Analysis Report 3. Plan for Week 3 and Week 4: Project 1 Stage 3- Analyze photos and videos of CCTV inspection in pipelines	1. Stakeholder Analysis Report completion 2. Plan for Stage 3-Analyze photos and videos of CCTV inspection in pipelines	1. Realize there are some difficulties to get the CCTV inspection in pipeline photos and videos data, we start to discuss how to solve this problem 2. We plan to find datasets for Stage 3 and postpone Stage 2 Competitor Analysis Report after Stage 3 3. Discuss use cases with Mehul to see if we can solve the datasets problem
Week 3: 6/22-6/26	1. Contact external stakeholders such as Hades AI(a company using innovative way to inspect pipelines) and CMU professors to ask if they are willing and able to provide CCTV in pipeline data	1. Project 1 Stage 2- Competitor Analysis Report Completion	1. Unable to find anyone to provide us CCTV inspection dataset a. Meet with Hades AI but fail: reluctant to tell us any industrial advice b. Contact CMU professors but fail: no

	2. Re-plan and continue completing Project 1 Stage 2- Competitor Analysis Report		one has the experience in the similar field 2. Complete Project 1 Stage 2- Competitor Analysis Report Completion
Week 4: 6/29-7/3	1. Discuss with our project manager and advisor about the feasibility of changing a new topic 2. Find alternative topics and datasets to propose with TCS	1. Alternative Project topic decision- Electricity Consumption Forecast(Project 2) 2. Datasets finding for adding attributes to forecast electricity consumption	1. Decide our topic as Electricity consumption forecast in U.S 2. Find at least ten attributes to merge with our original datasets found in EIA website.
Week 5: 7/6-7/10	1. Clean our datasets for future use 2. Complete EDA report 3. Plan for applying models to our dataset	1. Datasets Cleaning 2. Project 2- Exploratory Data Analysis(EDA) Report completion	1. Complete EDA report and present to Mehul 2. Get advice from Mehul and make revision
Week 6: 7/13-7/17	1. Complete Regression Analysis 2. Start ARIMA Analysis(7/17)	1. Complete Regression Analysis 2. Complete Artificial Neural Network Analysis	1. Conduct both countrywide and state-wise regression analysis 2. Decide to complete applying ANN model to our dataset
Week 7: 7/20-7/24	1. Complete Final report 2. Complete Final Presentation Slides 3. Rehearsal for final presentation	1. Final Report 2. Final Presentation slides	

- Internal Meeting time

Day	Time	Duration
Monday	4:30 p.m., 10:00 a.m.	1 hr
Tuesday	5:00 p.m.	30 min
Wednesday	3:30 p.m., 10:00 a.m.	1 hr
Thursday	5:00 p.m.	30 min
Friday	4:00 p.m.	1 hr

- External Meeting time

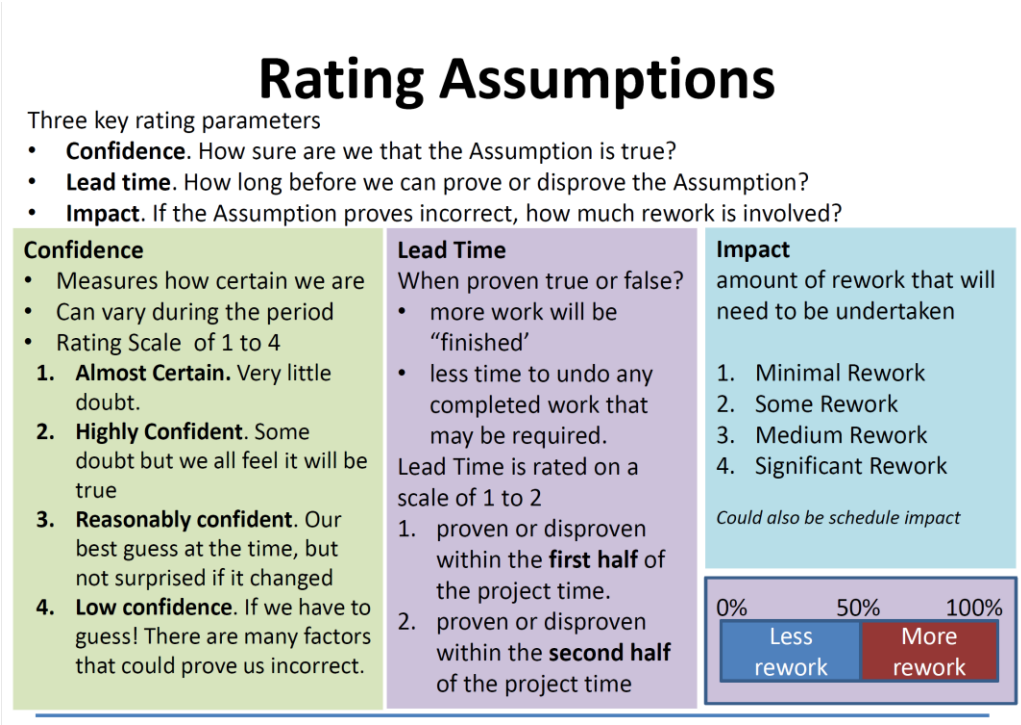
Day	Time	Duration
Thursday	3:00 p.m.	1 hr
Friday	4:00 p.m.	1 hr

- Advisor Meeting time

Day	Time	Duration
Wednesday	2:30 p.m.	1 hr

3. Assumptions, Dependencies & Constraints

3.1 Assumptions



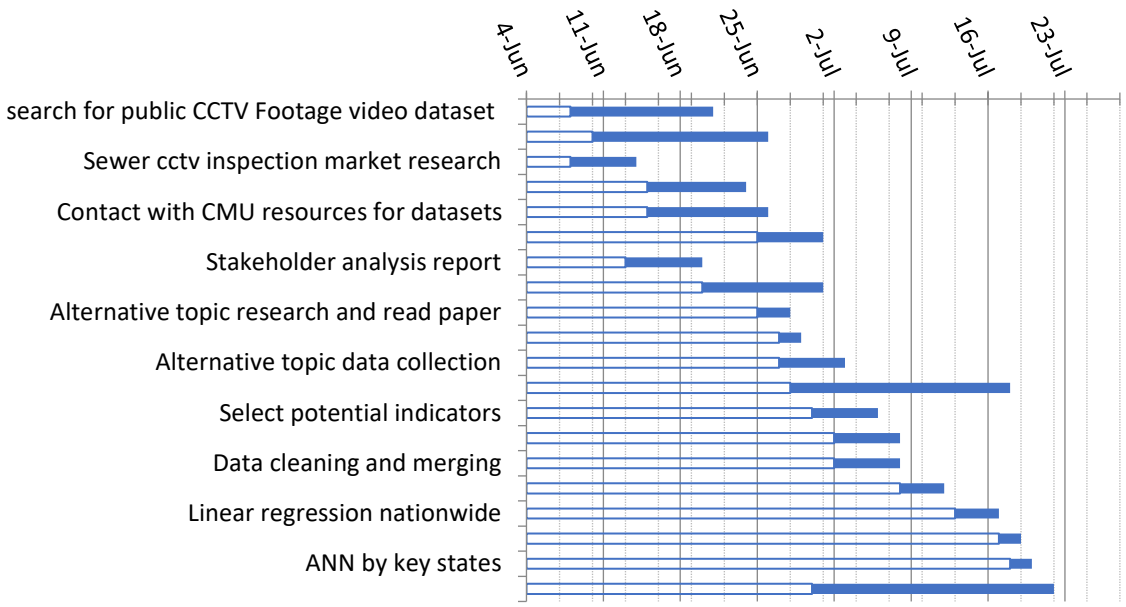
Source: The Northern Nevada Chapter

Time	Assumption	Confidence	Lead Time	Impact	Total Score
Week1	We are able access data from TCS	Low confidence (4)	First half (1)	Significant Rework (4)	9
Week1	We could learn computer vision techniques and image classification	Low confidence (4)	First half (1)	Significant Rework (4)	9
Week1	All relevant stakeholders will come to the next meeting.	Highly Confident (2)	Second half (2)	Minimal Rework(1)	3
Week 2	Data could be fetched from Heinz research resources	Reasonably Confident (3)	First half (1)	Significant Rework (4)	8
Week2	Data could be fetched from private company	Low confidence (4)	First half (1)	Significant Rework (4)	9

Week3	All team member are capable of doing static analytics	Almost Certain (1)	First half (1)	Minimal Rework (1)	3

3.2 Dependency

	Start	End	Duration
search for public CCTV Footage video dataset	8-Jun	19-Jun	13.0
computer vision literature review	10-Jun	25-Jun	16.0
Sewer cctv inspection market research	8-Jun	13-Jun	6.0
Contact with private company for insghts and data	15-Jun	23-Jun	9.0
Contact with CMU resources for datasets	15-Jun	25-Jun	11.0
Topic change	25-Jun	30-Jun	6.0
Stakeholder analysis report	13-Jun	19-Jun	7.0
Competitor analysis report	20-Jun	30-Jun	11.0
Alternative topic research and read paper	25-Jun	27-Jun	3.0
Alternative topic proposal	27-Jun	28-Jun	2.0
Alternative topic data collection	27-Jun	2-Jul	6.0
Electrcity consumption forecasting method research	28-Jun	17-Jul	20.0
Select potential indicators	30-Jun	5-Jul	6.0
Collect data	2-Jul	7-Jul	6.0
Data cleaning and merging	2-Jul	7-Jul	6.0
EDA	8-Jul	11-Jul	4.0
Linear regression nationwide	13-Jul	16-Jul	4.0
Linear regression by key states	17-Jul	18-Jul	2.0
ANN by key states	18-Jul	19-Jul	2.0
Final report	30-Jun	21-Jul	22.0



3.3 Constraints

ID	Constraints	Active
1	CCTV footage Dataset are unavailable	Yes
2	Limited Information we are able to obtain in this new area	Yes
3	The granularity of electricity consumption data by states are monthly data.	Yes
4	The GDP data are only in annual scale	Yes
5	Electricity vehicle data are in annual scale	Yes
6	No humidity data by state by month found for US	Yes
7	Limited time to apply new algorithms such as ARIMA and LSTM	Yes

4. Project 1 Sewer Wastewater Market Research

4.1 Project Description

Water and Wastewater Utilities across the world maintain and manage sewer and storm water pipeline network to ensure it provide hygienic and safe conditions for the citizens to live. This pipeline network requires regular inspection and maintenance. Most of this infrastructure is very old (setup decades and even century ago). Utilities spend millions of dollars annually for this purpose. One of the ways in which the Utilities carry out this activity is sending a CCTV camera inside the pipeline and capture the video footage and then manually go through them to identify issues like cracks, blockages, damages etc.

Utilities typically hire contractors for this job to capture the video footages and then have them go through manually to identify issues. There is an opportunity to create AI/Computer Vision based system to automate the process of inspecting these CCTV camera feed.

4.2 Project Objectives

Utilities industry is an Asset heavy industry with most of its asset base very old and towards end of the initial planned life. This asset base needs to be regularly inspected to ensure that potential issues are identified before the fault/issue arises and preventive actions taken. TCS' Utilities business is investing in this space to create Digital technology (e.g. AI, Computer Vision etc.) led solutions to help Utilities manage their Inspection and Monitoring processes more effectively.

The goal of our team is to help our client discover the market value of such AI/DL solutions, and also conduct research in terms of how the AI/Computer Vision solutions could be applied into such domain, how leading companies in this market work, and how the present and future market landscape looks like. We tried our best to approach an innovative and automatic way of using CCTV to maintain and inspect the sewer system, so to provide valuable insights for our clients.

4.3 Project Deliverables

4.3.1 Stakeholder Analysis on CCTV inspection in Sewage Industry

This report has been released in the progress of the project. Please refer to the following link:

Project 1 Deliverable-Stakeholder Analysis Report

4.3.2 Competitor Analysis on CCTV inspection in Sewage Industry

This report has been released in the progress of the project. Please refer to the following link:
Project 1 Deliverable-Competitor Analysis Report

1. Project 2 Energy Consumption Analysis

5.1 Project Description

In the second part, we will shift our focus to another utility related issue: the electricity consumption in the United States. In our analysis, we will be using the retail sales of electricity to represent the electricity consumption in USA. The reason for this is that retail sales of electricity constitutes 96% of the total consumption of electricity [1].

For the second project, we will explore different factors, and how they affect the consumption of electricity. For example, in some places, when the season is winter or summer, electricity consumption will be higher than spring and autumn because of cold or hot weather. Or when there are more people in an area, there will be more electricity consumption.

We will first demonstrate different factors that we choose to understand the consumption of electricity, and how they are individually related to the overall consumption. Then we will implement regression on these factors to make predictions on future electricity consumptions, so as to provide insight for our clients.

[1] US energy information administration, https://www.eia.gov/energyexplained/electricity/data-and-statistics.php, Nov 14, 2019

5.2 Project Objectives

We want to first identify factors that are of vital importance to help understand the consumption of electricity, and we will do exploratory data analysis to demonstrate their relationship. Then we will do regression analysis on a nation-wide scale as well as a state wise scale, so as to understand which factors to look into when we want to study and estimate the electricity consumption.

5.3 Project Deliverables: Energy Consumption Forecast & Time Series Analysis

5.3.1 Dataset Description

Our Electricity Consumption data is merged from several datasets, including:

- Electricity Retail Sales: State-wise, 3 sectors(Residential, Industrial, Commercial)
- Electricity Retail Price: State-wise, 3 sectors(Residential, Industrial, Commercial)
- Weather Dataset: station-wise, temperature snow and precipitation data averaged by states

- Population
- State size :Total land and water area surveyed in 2018
- Solar generation: state-wise generation of electricity by distributed solar photovoltaic(Residential, Industrial, Commercial)

We combined the datasets into a merged dataset. The features are as the followings:

Feature Name	Description	Unit/Value range
Year	Year data from 2014-2019	year
Month	12 months data included in every year	month
AK~WY	50 dummy variables representing State	1 or 0
CLDD	Cooling Degree Days - Computed when daily average temperature is above 18.3°C/65°F. CDD = mean daily temperature -18.3° to tenths degree Celsius. Each day is summed to produce a monthly total.	ºF
TAVG	Average Monthly Temperature - computed by adding the unrounded monthly mean TMAX (average of the daily maximum temperatures) and TMIN temps (average of the daily minimum temperatures) and dividing by 2; then round to hundredths degree Celsius. Values are set to missing if either the monthly mean TMAX or TMIN temperature is missing.	ºF
AWND	Monthly (Annual) average wind speed. Average the Daily AWND values in GHCN-D to get monthly and annual averages. (tenths of meters per second).	tenths of meters per second
HTDD	Heating Degree Days - computed when daily average temperature is less than 18.3°C/65°F. HDD = 18.3° – mean daily temperature to tenths degree Celsius.	ºF

	Each day is summed to produce a monthly total.	
area	Total land and water area of the United States by state and territory	Square miles
population	State-level monthly data are linearly estimated through available yearly data	person
solar-generation	Electricity generated by distributed solar-PV	mkwh
Spring, Summer, Fall, Winter	4 dummy variables representing seasons March, April, May: Spring June, July, August: Summer September, October, November: Fall December, January, February: Winter	1 or 0
Residential_Retail Price	The monthly average residential electricity retail price of each State	Cents/kWh(kilowatt-hour)
Industrial_Retail Price	The monthly average industrial electricity retail price of each State	Cents/kWh(kilowatt-hour)
Commercial_Retail Price	The monthly average commercial electricity retail price of each State	Cents/kWh(kilowatt-hour)
Residential_Usage	Retail sales of electricity in the residential sector	mkwh(million kilowatt-hour)
Industriall_Usage	Retail sales of electricity in the industrial sector	mkwh(million kilowatt-hour)
Commercia_Usage	Retail sales of electricity in the commercial sector	mkwh(million kilowatt-hour)

5.3.2 Exploratory Data Analysis(EDA)

1. How the United States uses energy

Residential Sector: includes homes and apartments.

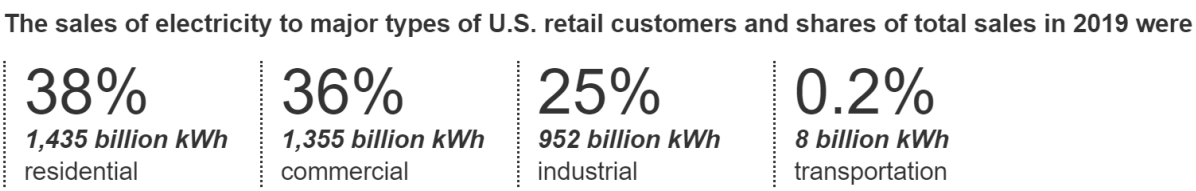
Commercial Sector: includes offices, malls, stores, schools, hospitals, hotels, warehouses, restaurants, and places of worship and public assembly.

Industrial Sector: includes facilities and equipment used for manufacturing, agriculture, mining, and construction.

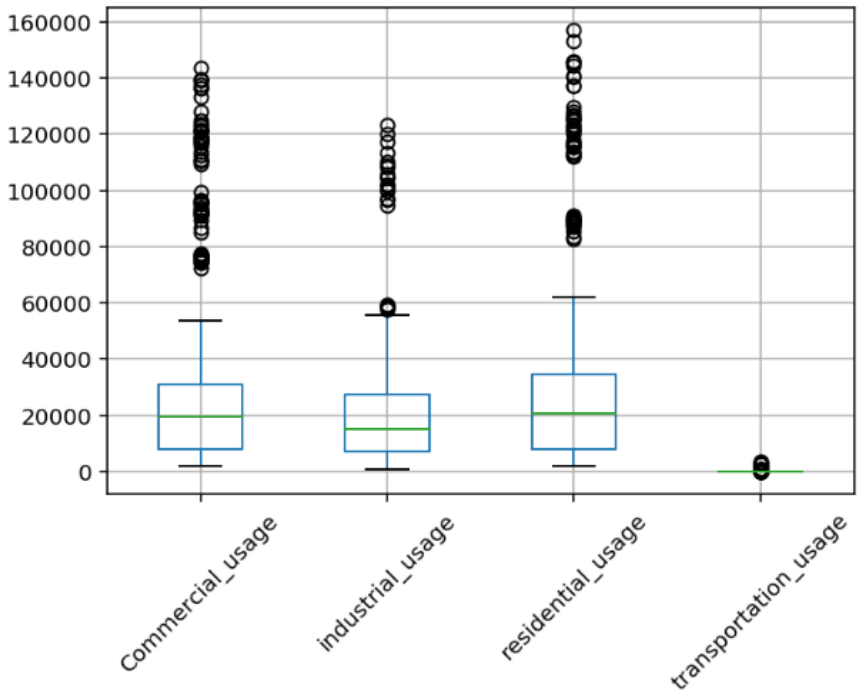
Transportation Sector: includes vehicles that transport people or goods, such as cars, trucks, buses, motorcycles, trains, aircraft, boats, barges, and ships.

2. Retail sales of electricity

Retail sales include net imports (imports minus exports) of electricity from Canada and Mexico.



https://www.eia.gov/energyexplained/electricity/electricity-in-the-us-generation-capacity-and-sales.php



Transportation usage accounts only a small portion of electricity usage among all sectors, so we only focus the rest three sectors.

Every year, residential use of electricity among states takes the largest portion of total sales and also has the greatest variance among all sectors.

3. Explore factors

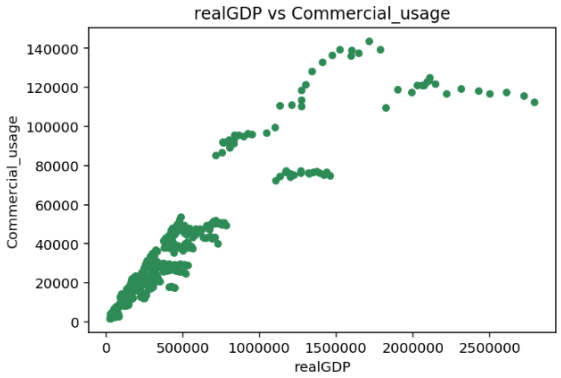
1. GDP and Population

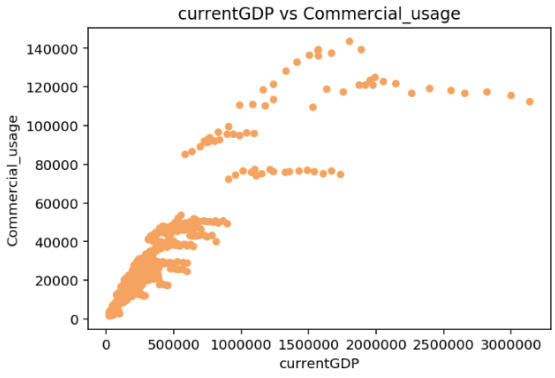
	currentGDP	realGDP	GDP-per-capita(current-dollars)	GDP-per-capita(chained-2012-dollars)	Population	Commercial_usage	industrial_usage	residential_usage	transportation_usage
currentGDP	1.00	0.99	0.30	0.27	0.97	0.91	0.61	0.79	0.58
realGDP	0.99	1.00	0.26	0.26	0.98	0.93	0.63	0.80	0.59
GDP-per-capita(current-dollars)	0.30	0.26	1.00	0.91	0.15	0.12	-0.08	0.00	0.31
GDP-per-ita(chained-2012-dollars)	0.27	0.26	0.91	1.00	0.15	0.11	-0.09	-0.02	0.33
Population	0.97	0.98	0.15	0.15	1.00	0.97	0.68	0.88	0.51
ommercial_usage	0.91	0.93	0.12	0.11	0.97	1.00	0.72	0.95	0.46
industrial_usage	0.61	0.63	-0.08	-0.09	0.68	0.72	1.00	0.77	0.12
residential_usage	0.79	0.80	0.00	-0.02	0.88	0.95	0.77	1.00	0.28
sportation_usage	0.58	0.59	0.31	0.33	0.51	0.46	0.12	0.28	1.00

From the factor perspectives, Population has the highest correlation value in almost all sectors compared to current GDP, real GDP, and GDP per capita. From the electricity usage sector point, The correlation coefficients to all the attributes ranks from high to low are respectively commercial, residential, industrial and transportation. It's not hard to explain considering commercial and residential activity are stimulated by economic performance and are concentrated on big cities. On the other hand, and transportation consumption and industrial consumption though also positively correlated with GDP and population, are likely to be associate with countries and farming driven economy.

GDP

Commercial

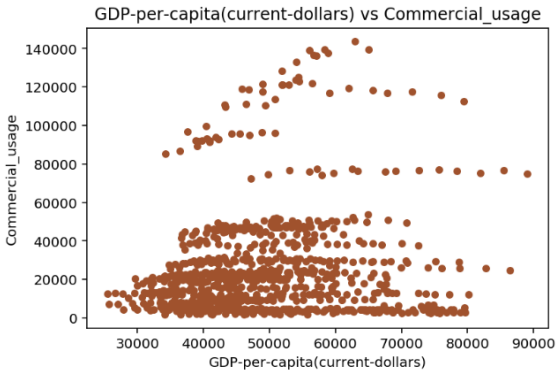


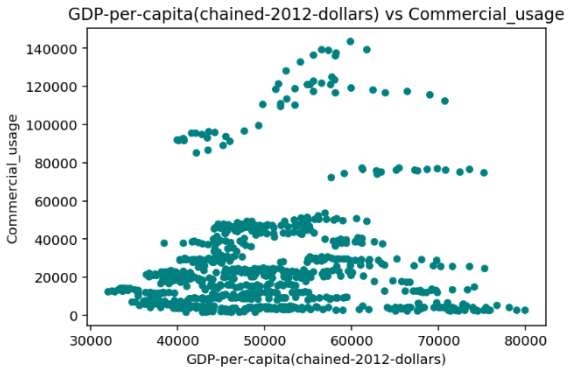


Based on the scatter plot, there seems to be a positive correlation between GDP and commercial consumption.

When current and real GDP are below around 800,000 millions of dollars, GDP is strongly correlated to commercial consumption

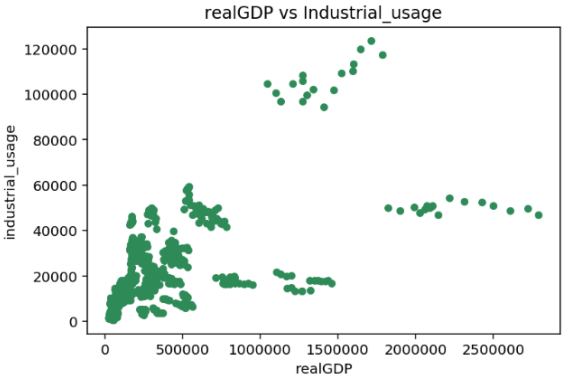
When current and real GDP are above 500,000 millions of dollars, there are <u>three groups of industrial demand trend among states</u>

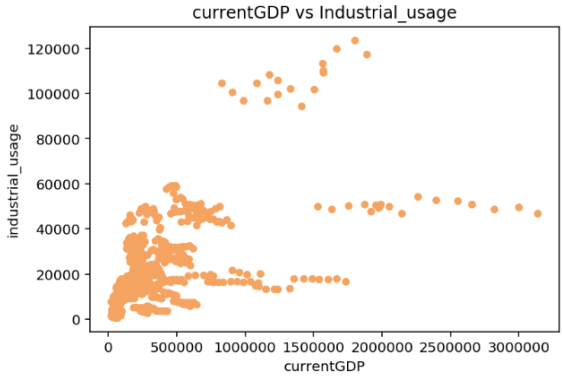




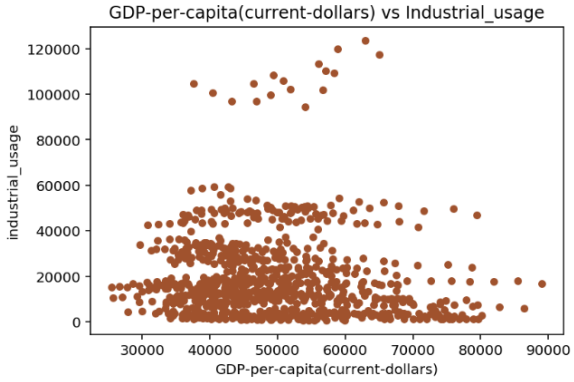
Based on the scatter plot, there seems to be no relationship between GDP per capita and commercial usage.

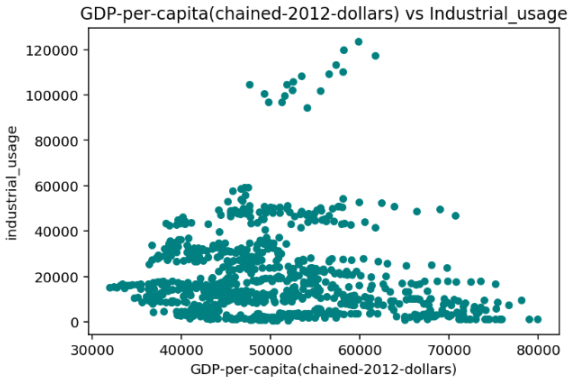
Industrial





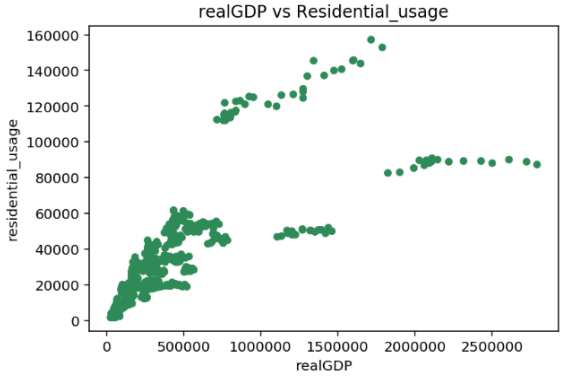
When current and real GDP are above 500,000 millions of dollars, there are <u>two groups of industrial demand trend among states</u>

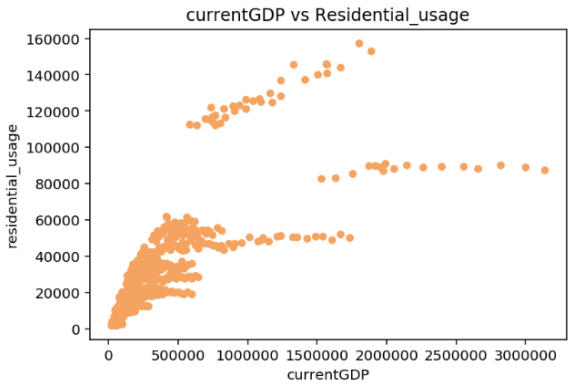




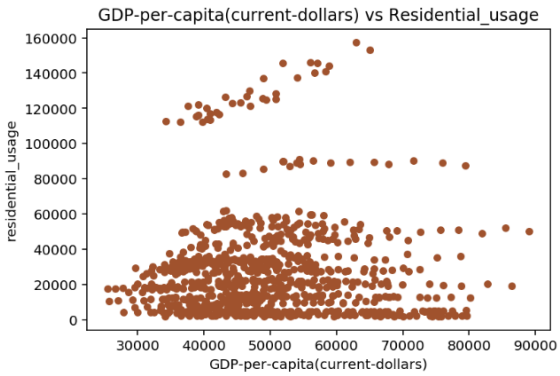
Based on the scatter plot, there seems to be no relationship between GDP per capita and industrial usage.

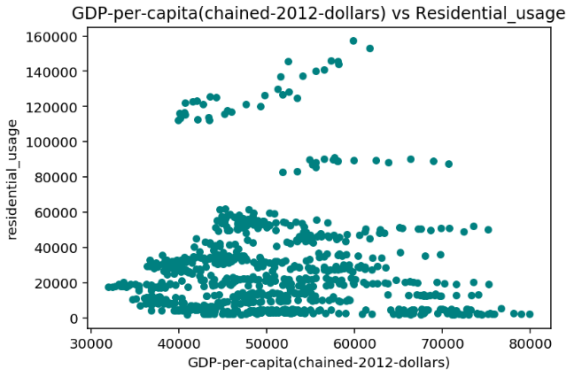
Residential





When current and real GDP are above 800,000 millions of dollars, there are <u>two groups of residential demand trend among states</u>

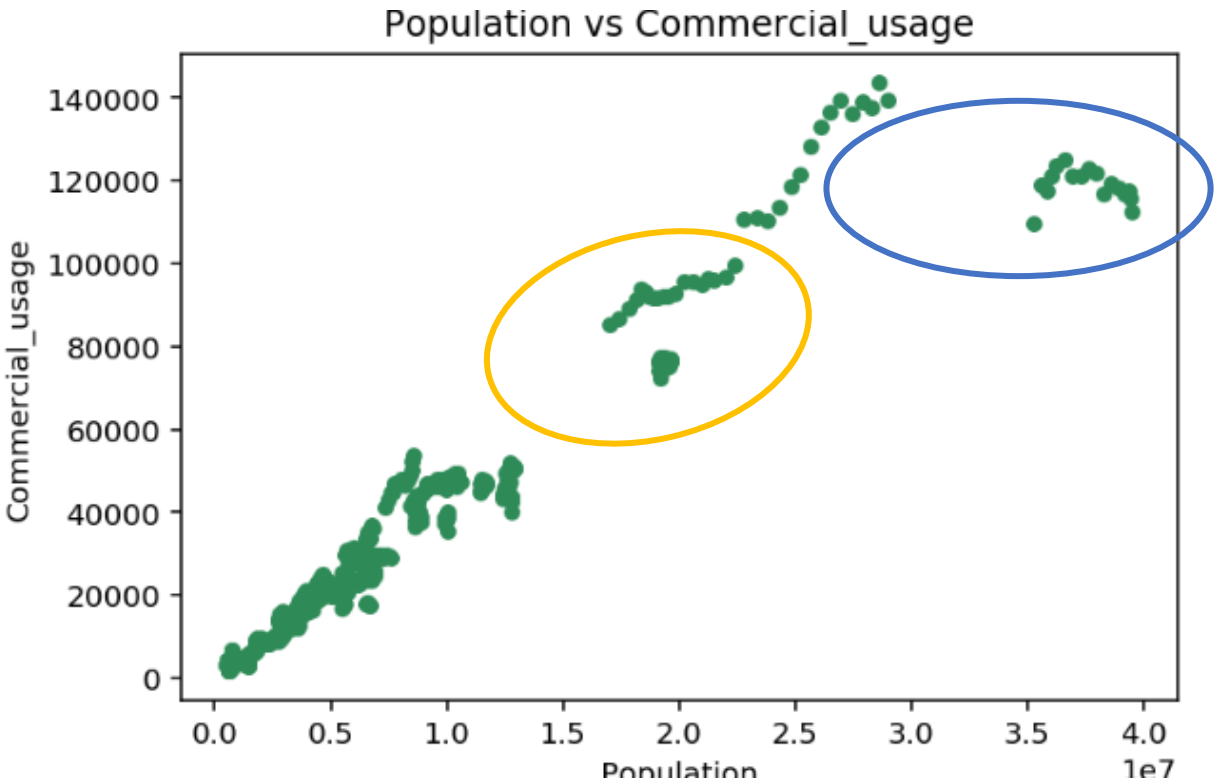




Based on the scatter plot, there seems to be no relationship between GDP per capita and residential usage.

Population

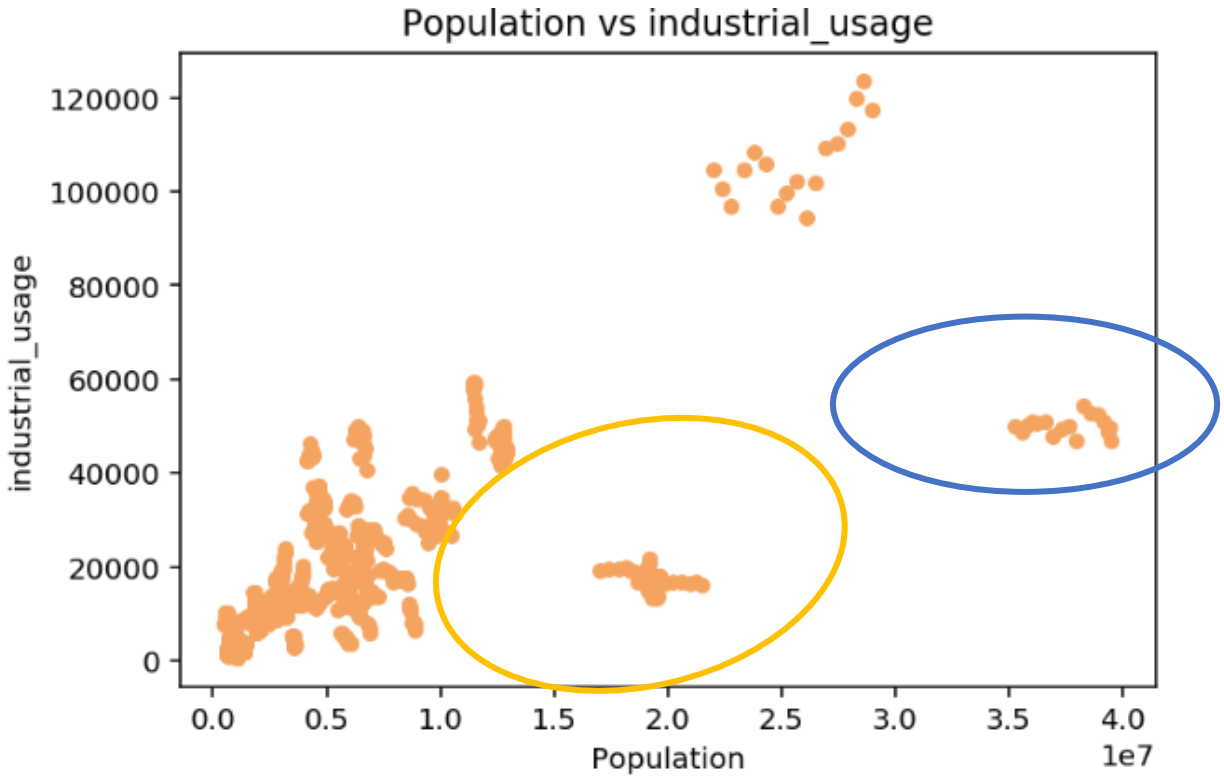
Commercial



Based on the scatter plot, there seems to be a medium to strong correlation between population and commercial consumption.

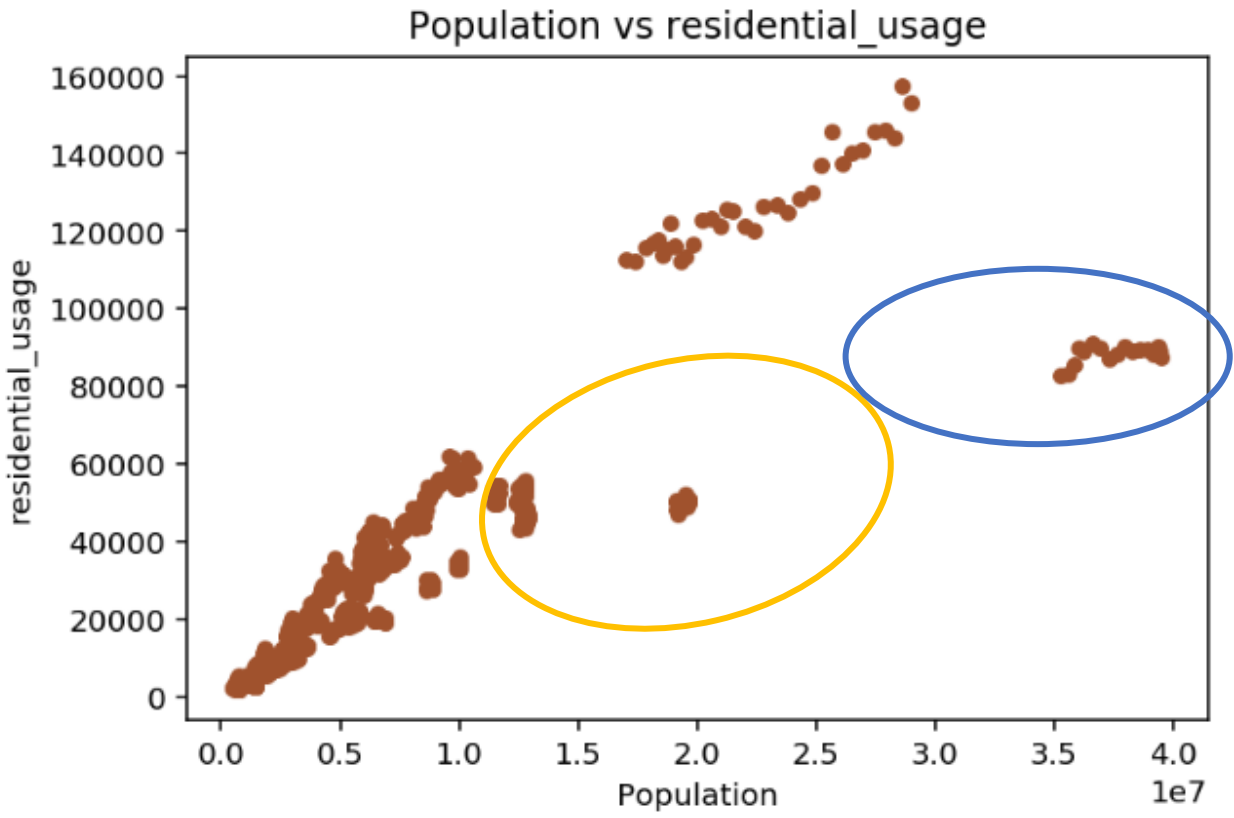
Between population range from 1.5 to 2.5 million and 3.5 to 4.0 million, the correlation is not positive. SO we investigate into these points and found the yellow circled points are data points from New York and Florida, while blue circled points are from California.

Industrial



Same with commercial usage, NY, FL and CA are also three states fall into the two circles in terms of industrial consumption. These are all states with high density population and less industrial constructions.

Residential



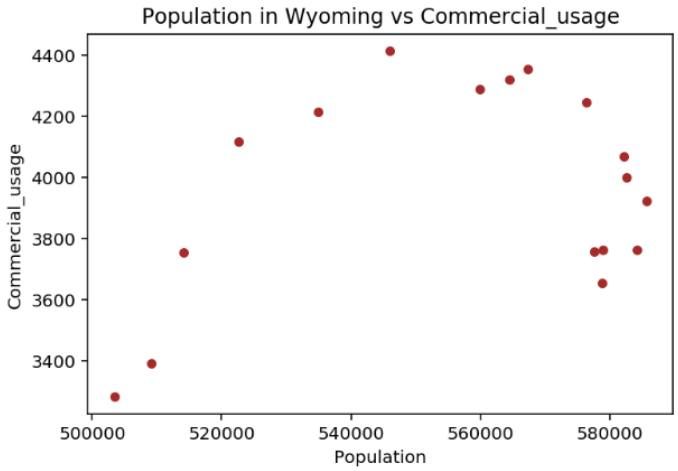
Similar to the Commercial usage scatter plot, when the state population is below 10 million people, number of people seems to be strongly correlated to commercial consumption. The yellow and blue line encircle New York and California while Florida goes up to align with the regression line.

State-example

based on the graph when the state population is below 10 million people, number of people is highly correlated to commercial consumption. Thus I tried to explore the detailed relationship by selecting one state that the population is below 10 million people.

Wyoming has lowest population in the united states, and the number of people in this state is below 1 million people. Thus, I Select the state Wyoming to explore the relation

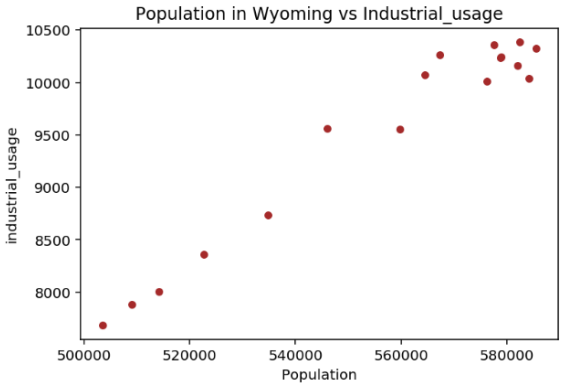
After specifically choosing a state where the population is below 10 million men and women, the connection seems less apparent



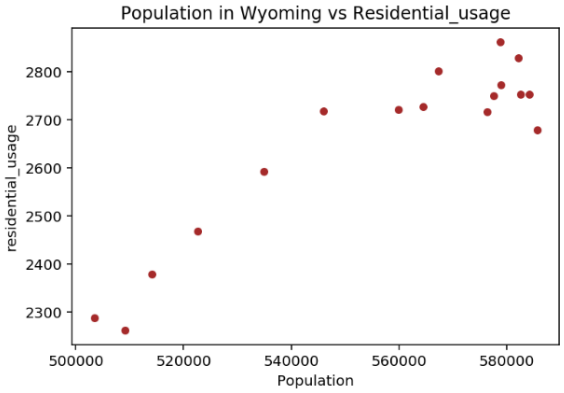
	Population	Commercial_usage
Population	1.00	0.33
Commercial_usage	0.33	1.00

Average population of Wyoming in 2003-2019: 556942 people

The connection seems less apparent between population and commercial usage



	Population	industrial_usage
Population	1.00	0.98
industrial_usage	0.98	1.00

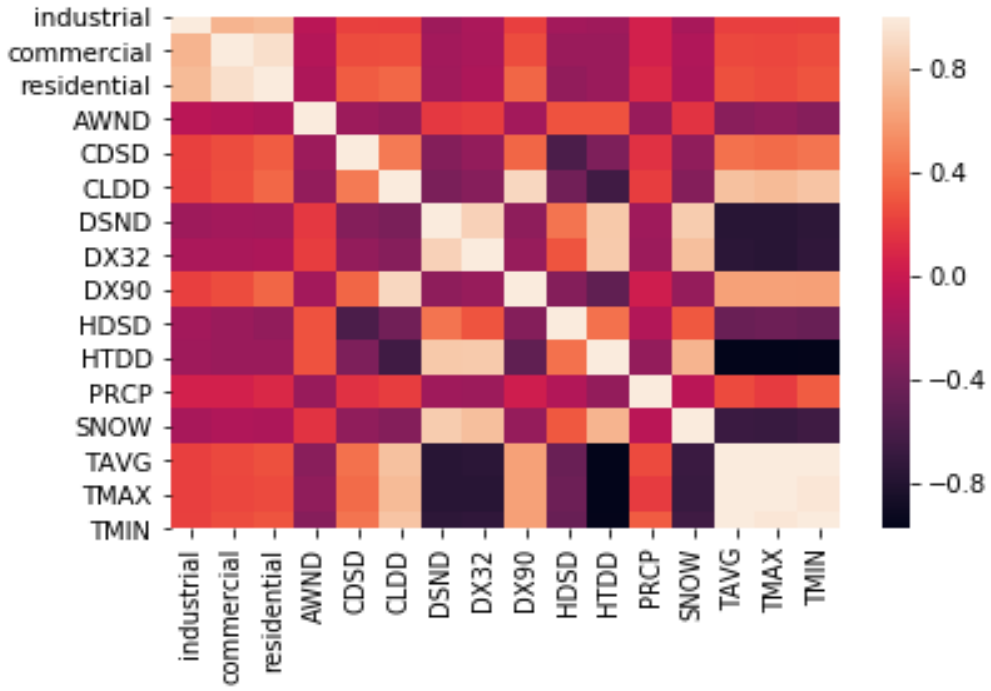


	Population	residential_usage
Population	1.00	0.92
residential_usage	0.92	1.00

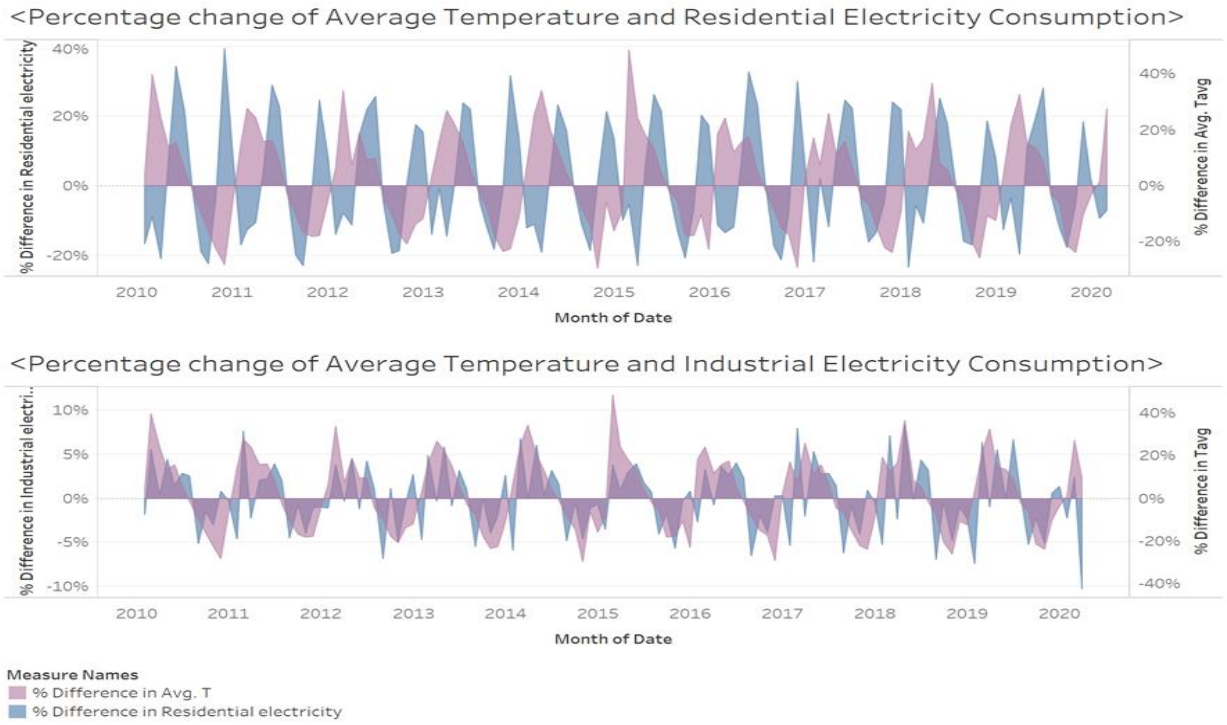
There seems to be a strong connection between population industrial and residential usage.

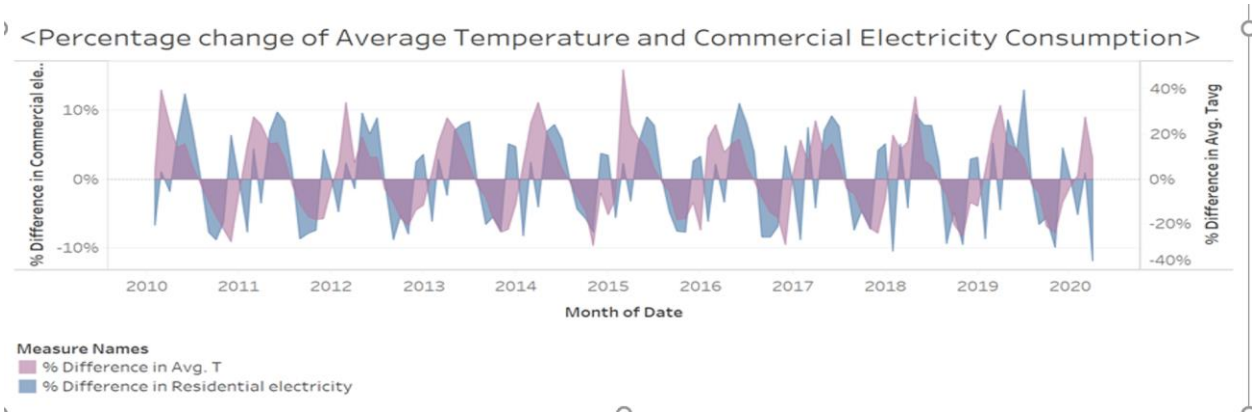
Climate

- **Snow**
    - SNOW - Snowfall
    - DSND - Number days with snow depth > 1 inch(25.4mm) for the period.
- **Precipitation**
    - **PRCP** - Precipitation
- **Temperature measured by day**
    - HTDD - Heating degree days
    - CLDD - Cooling Degree Days
    - HDSD - Heating Degree Days Season to Date
    - CDSD - Cooling Degree Days Season to Date
    - DX32 - Number days with maximum temperature < 32 F.
    - DX90 - Number days with maximum temperature > 90 F (32.2C)
- **Temperature measured by value**
    - TMAX - Maximum temperature
    - TAVG - Average Temperature.
    - TMIN - Minimum temperature
- **Wind**
    - AWND - Average wind speed



The correlation coefficient between Climate attributes and electricity consumptions are low. The highest one is between Cooling Degree Days and Residential electricity consumption with a value of 0.37.





The Percentage change of Temperature and ELectricity consumption have a clear seasonal trend. In Spring and Fall, The percentage change of average temperature rise to the highest and in summer and winter the percentage change return back to 0, matching the absolute highest and lowest temperature.

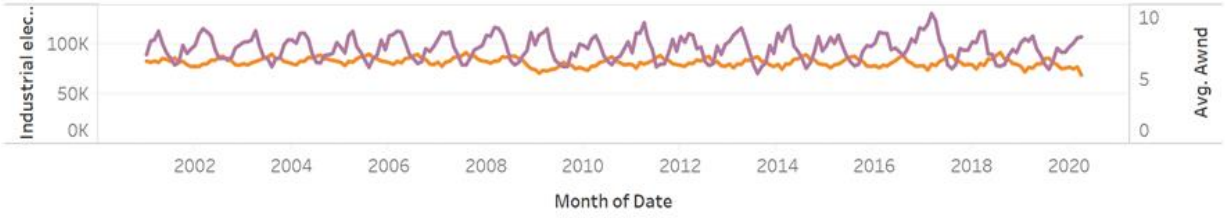
The Electricity trend goes in the opposite way. We could conclude that both high and low temperature would trigger an increase in electricity consumption.

when we compare consumptions from different sector, we could see residential consumption is more sensitive to temperature change than industrial one. Industrial consumption fluctuate in a monthly pattern rather than quarterly.

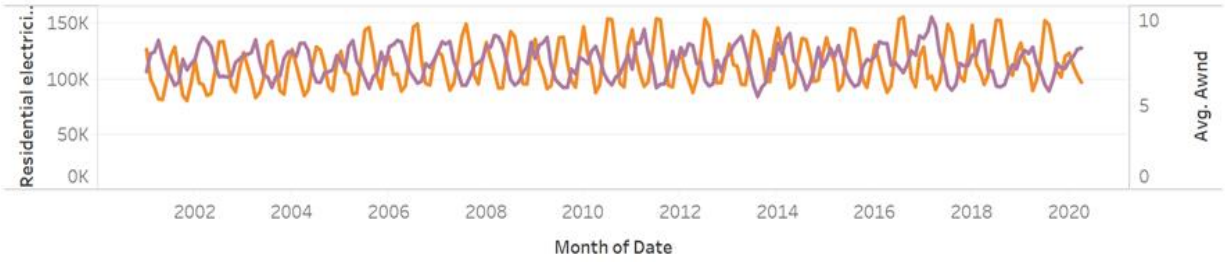
SImilarly when we compare consumptions from residential and commercial, we could see percentage change in commercial consumption is more like to residential.

The main reason for this difference across sectors could caused by difference in users.Electricity in Residential and Commercial sectors are closely related to people, while in industrial sector the main user of this energy are equipments and plants, thus are more independent with temperature change.

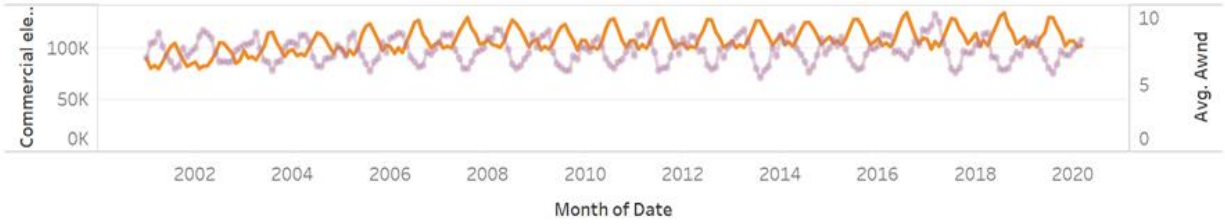
<Monthly Average Wind Speed & Industrial Electricity Consumption>



<Monthly Average Wind Speed & Residential Electricity Consumption>



<Monthly Average Wind Speed & Commercial Electricity Consumption>

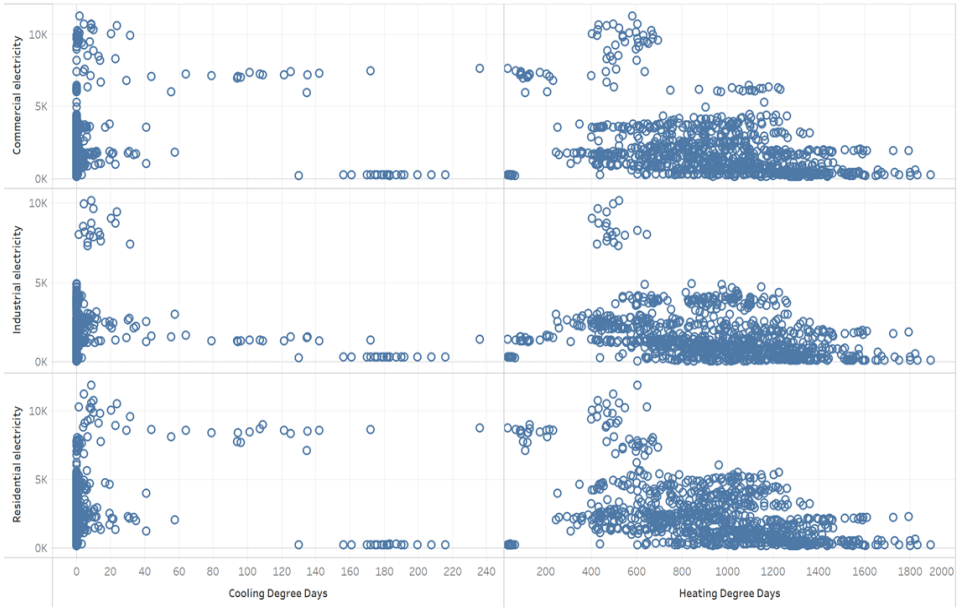


Likewise, the wind speed has a negative correlation with Electricity Consumption. And the Residential sector is most sensitive to the wind speed change.

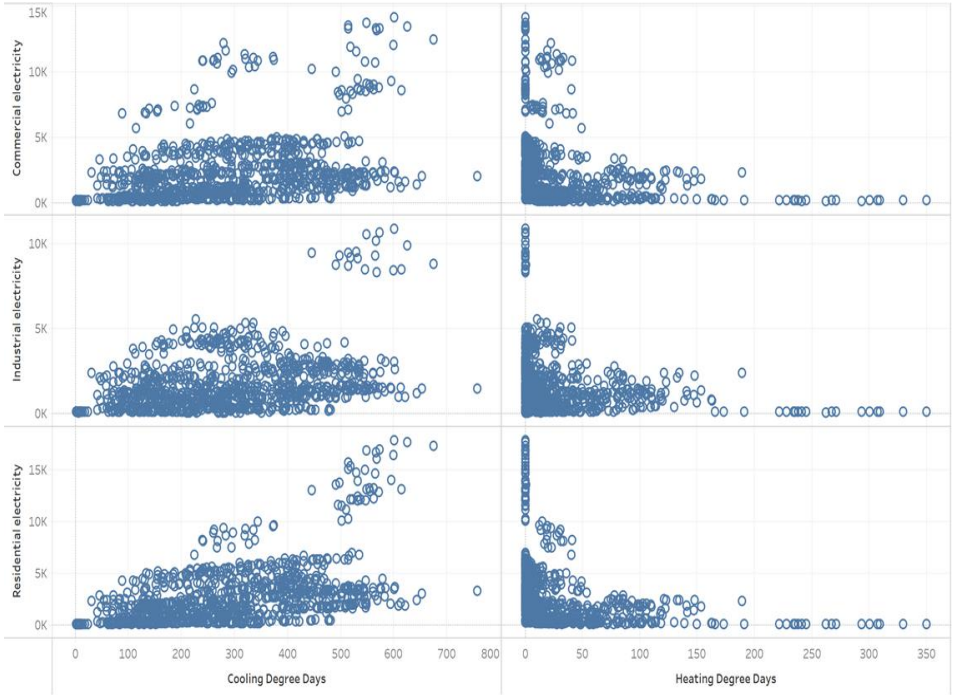
<Electricity Consumption and Monthly average wind speed>

Wind speed reach to the slowest in summer. No clear correlation is observed.

<Electricity Consumption & Heating Degree Days and Cooling Degree Days> - 12



<Electricity Consumption & Heating Degree Days and Cooling Degree Days> - 7



December:

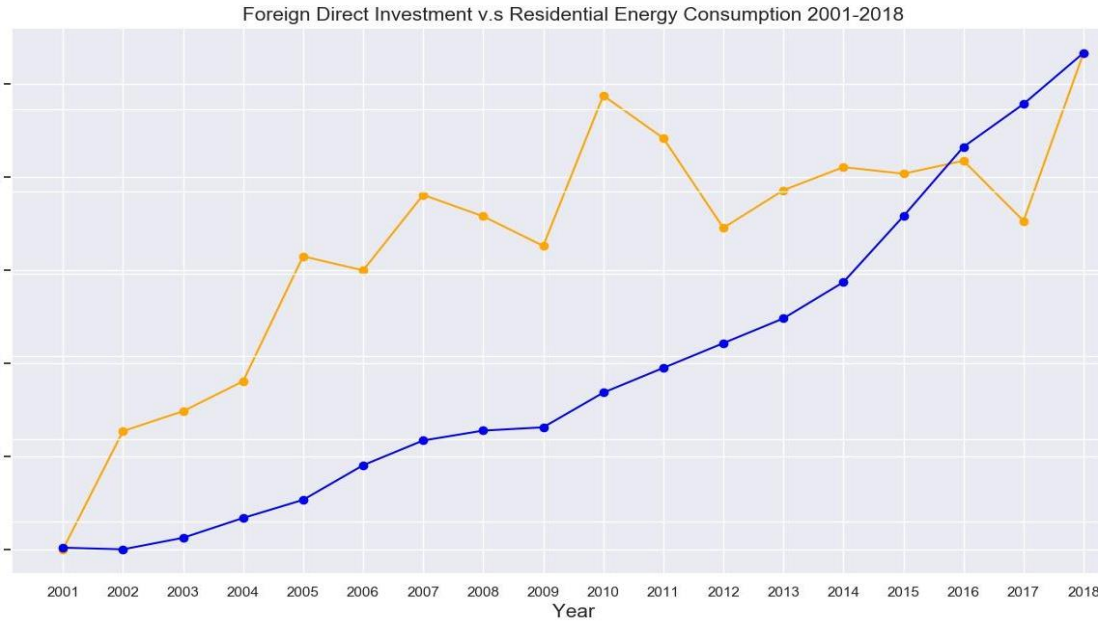
Cooling Degree Days decrease to the low point while Heating Degree Days achieve to the peak.

July:

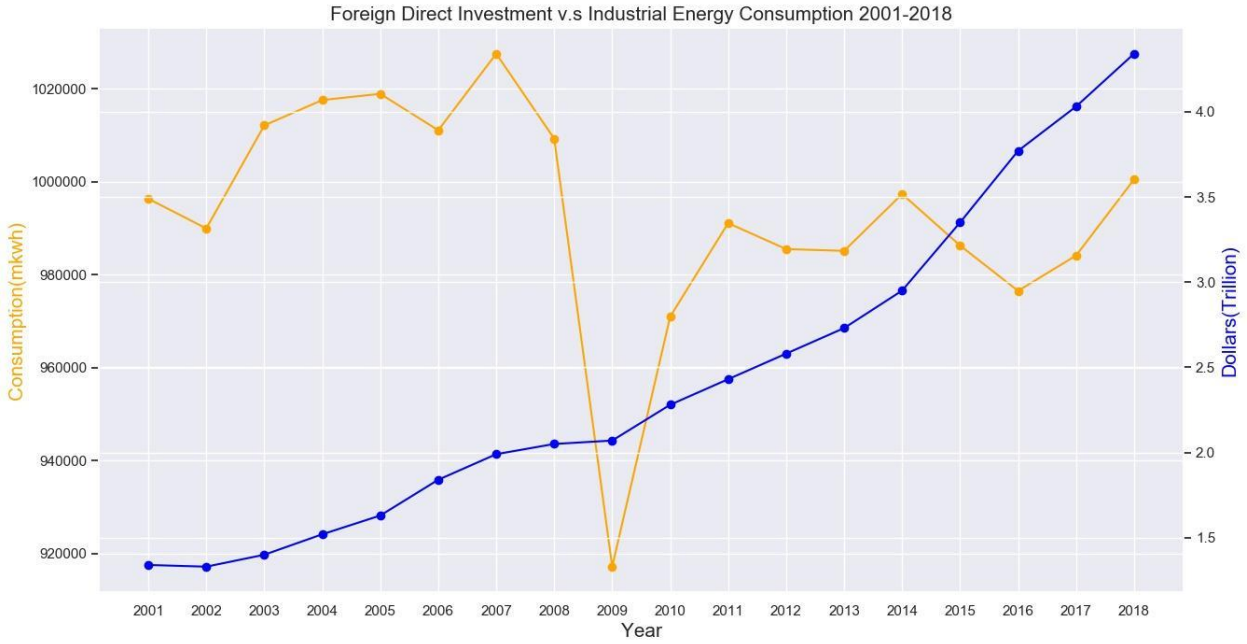
Cooling Degree Days and Heating Degree Days move to the opposite direction.

Note that when cooling degree days increases, electricity consumptions also increases.

Foreign Direct Investment



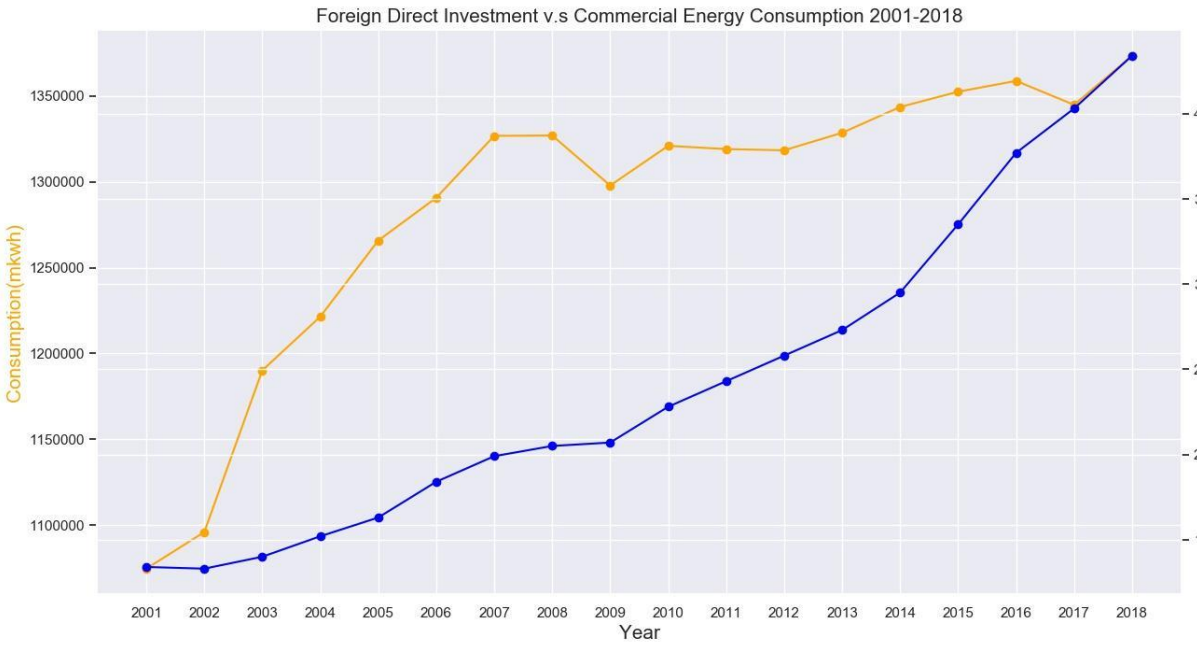
In general, residential electricity consumption is rising. However, it is not a smooth line and cannot see special reaction between FDI and residential energy consumption.



Unlike residential electricity consumption, industrial electricity consumption is not rising and generally remain at a level. However, in 2009, there's a big drop. It is an interesting topic to find out the reason and maybe that attribute can become a strong attribute to do analysis.

The reason for this drop is mainly global economic recession:

- According to The Norwegian American(2009), the electricity intensive industry is made up of industries with very high electricity consumption per produced unit.
- Their production is, to a large degree, **aimed at the export market**, which makes them more exposed to international economic cycles.
- The energy consumption in the **chemical and metal industry dropped in total by about 21 per** cent from 2008 to 2009, while the consumption in manufacture of **paper and paper products** fell by about **7 per cent**.

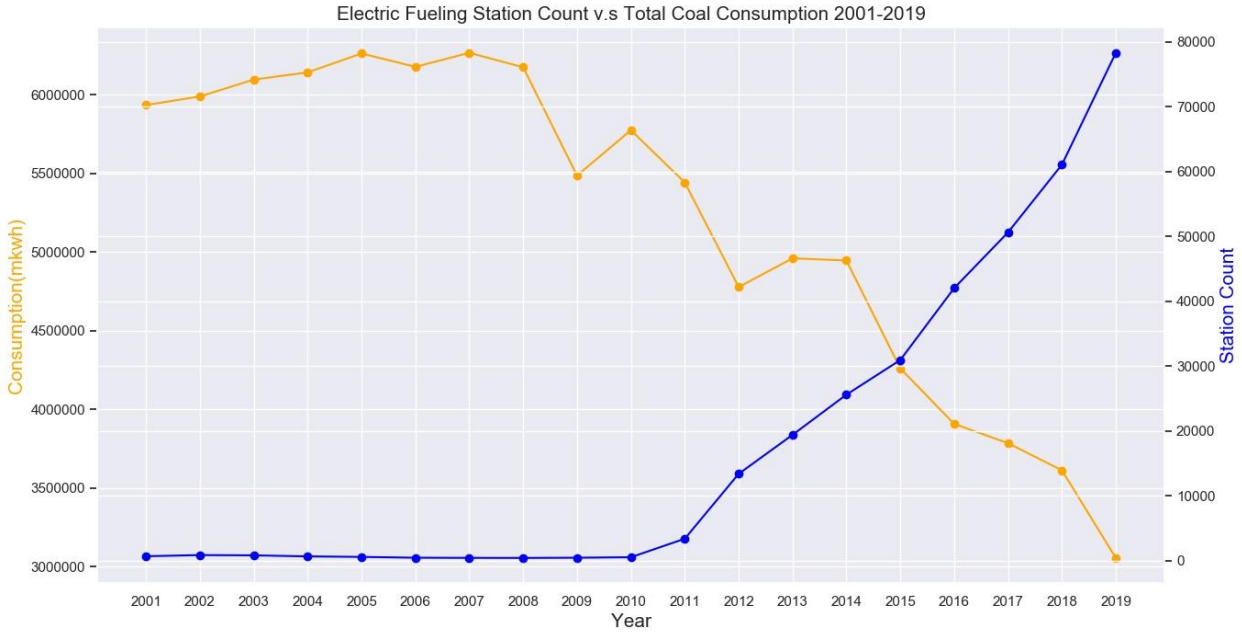


Compared to the previous two plots, commercial electricity consumption is rising smoothly and become stable in five years. There's also a small drop in electricity consumption in 2009 due to economic recession. But there's not specific indication show's that it is because of FDI.

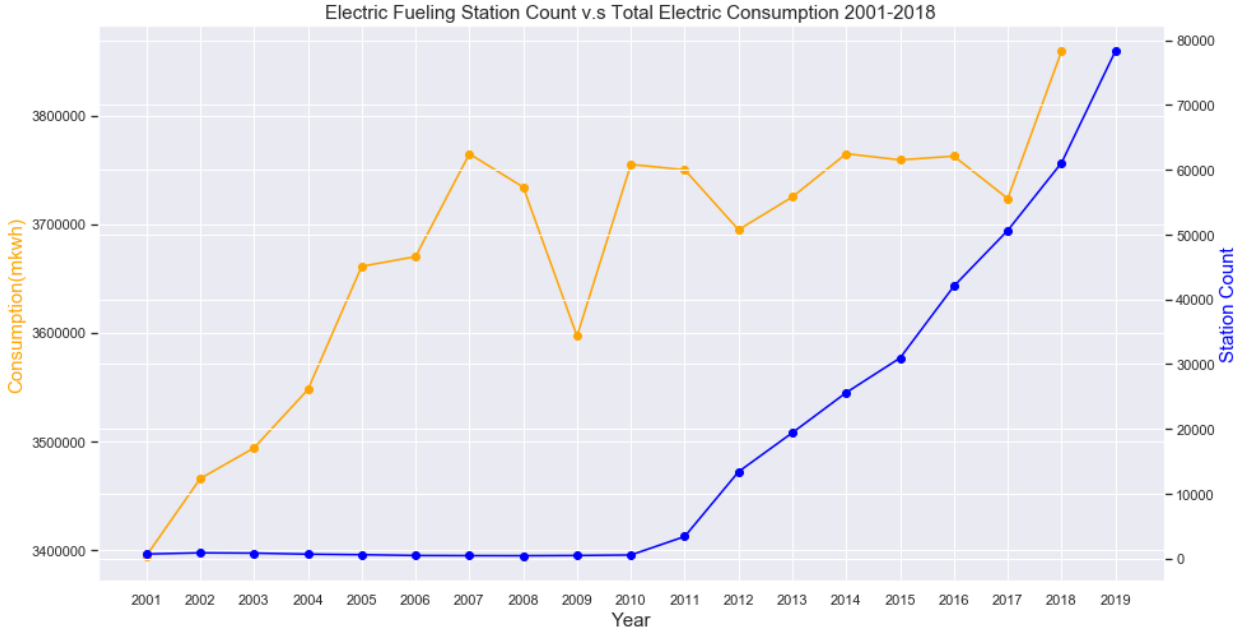
There's no specific pattern shown between electricity consumption and FDI in the previous plots. Therefore, this attribute might NOT be a strong attribute

when building models.

Fueling Charging Station



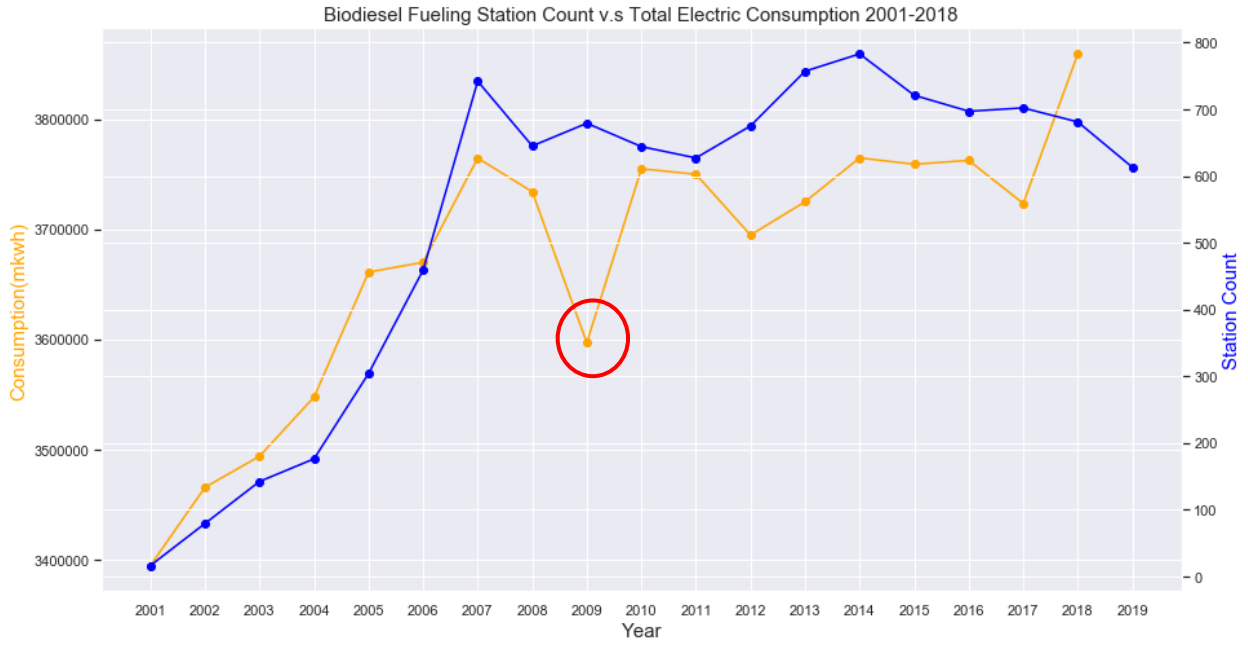
As expected, while electric fueling station count increases, total coal consumption drops.



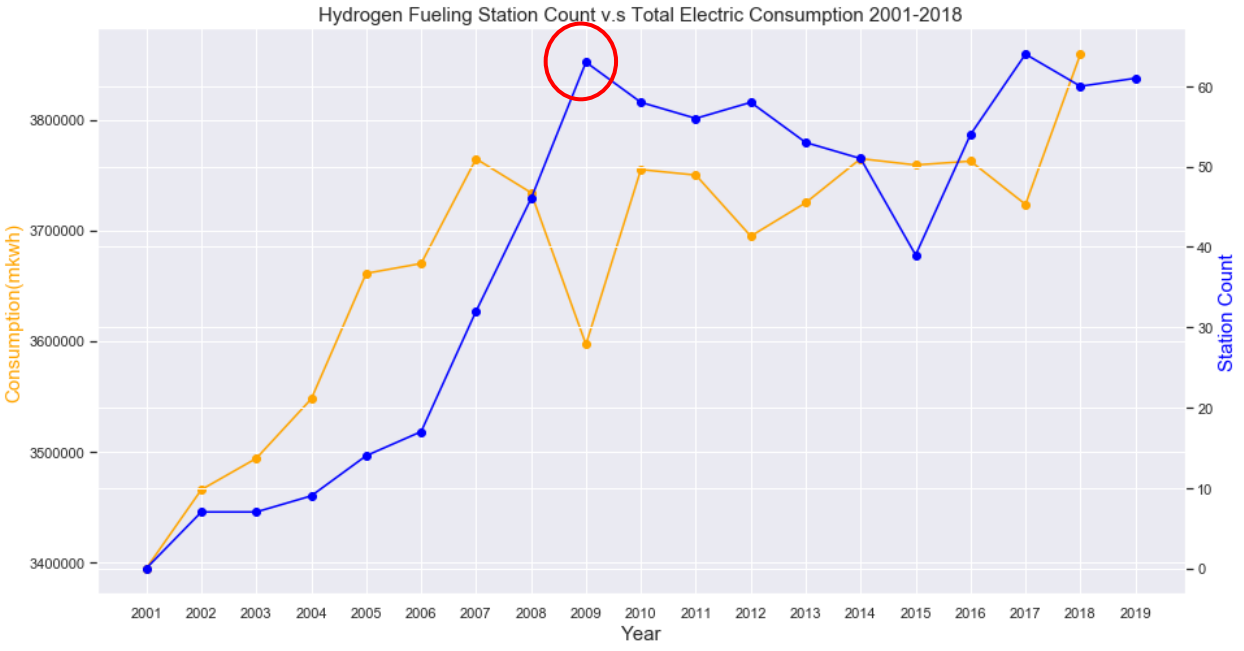
As electric fueling station count increases, total electric consumption increases, which is what we expected. Therefore, we can first assume there's

positive relationship between these two variables but we will do further analysis in the next step.

Biodiesel Fueling Charging Station Count vs. Total Electric Consumption



Hydrogen Fueling Charging Station Count vs. Total Electric Consumption



In the previous two plot, we can see even though electricity consumption dropped a lot in 2009, Biodiesel and Hydrogen fueling station count had risen.

- **This is especially due to lower consumption of gasoline in road transport and heavy fuel oil in shipping.** About 122 million litres of biodiesel were sold in 2009, which is 18 per cent less than the previous year. This made up about 4.8 per cent of diesel used for transport purposes in 2009.
- Some bioethanol was sold, but this makes up **only a tiny fraction** of the total sale of gasoline.
- Trade legislation was introduced for biofuels used in road transport, and the directed share rose from 2.5 per cent to 3.5 per cent from 1. April 2010.

Electricity Price

Most of the time series of the retail price of electricity in every state show rising trend

Most of the time series of the commercial and residential electricity consumption in every state shows stable trend in the past decades, especially in residential sector

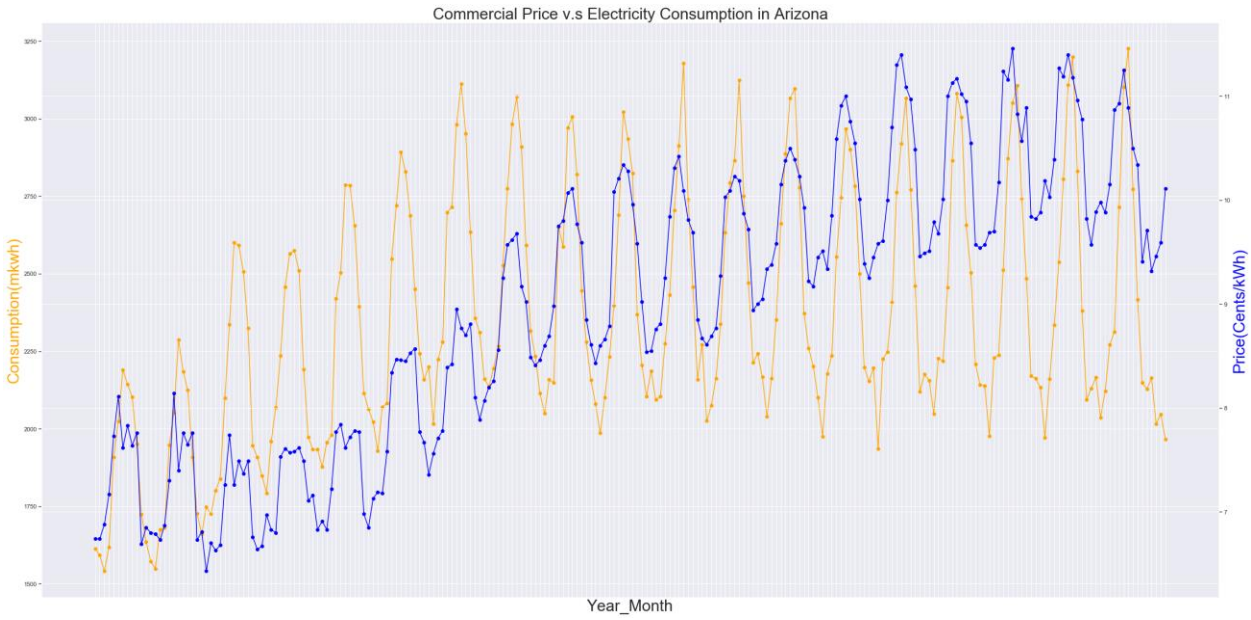
Industrial electricity consumption shows rising trend in some states but declines a lot in other states

Some of the state shows interesting trend that not similar to the others

Those interesting ones would show in the following slides. Interesting trends mostly appear in commercial and industrial sector.

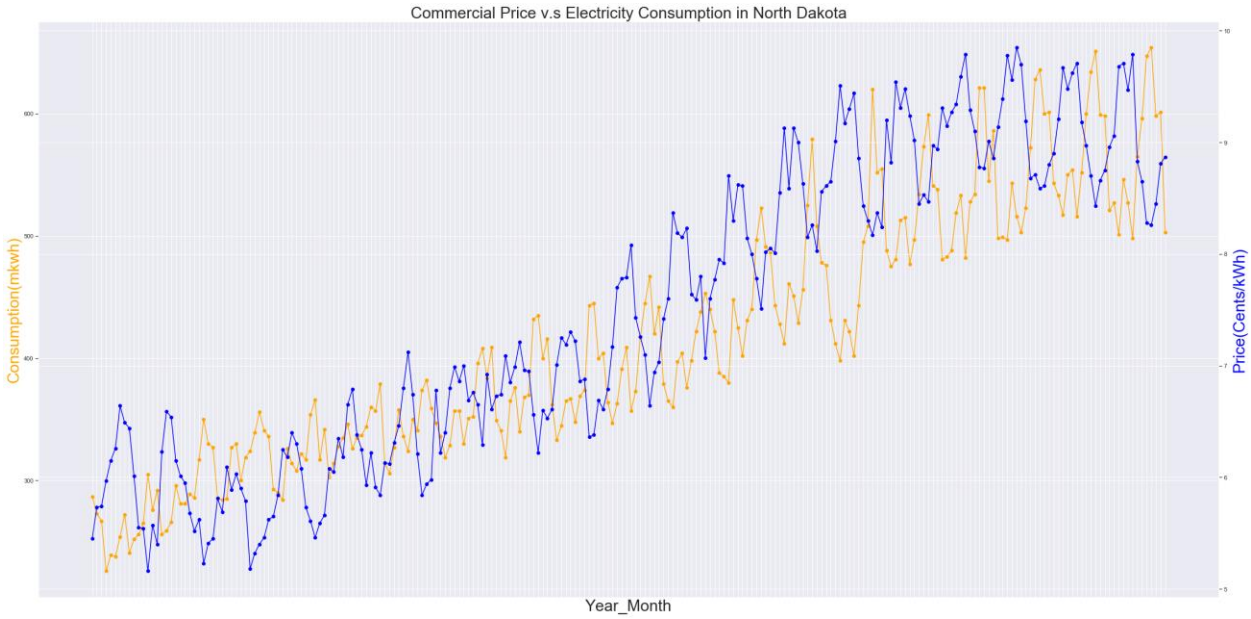
7-1. Commercial Price & Consumption

Arizona



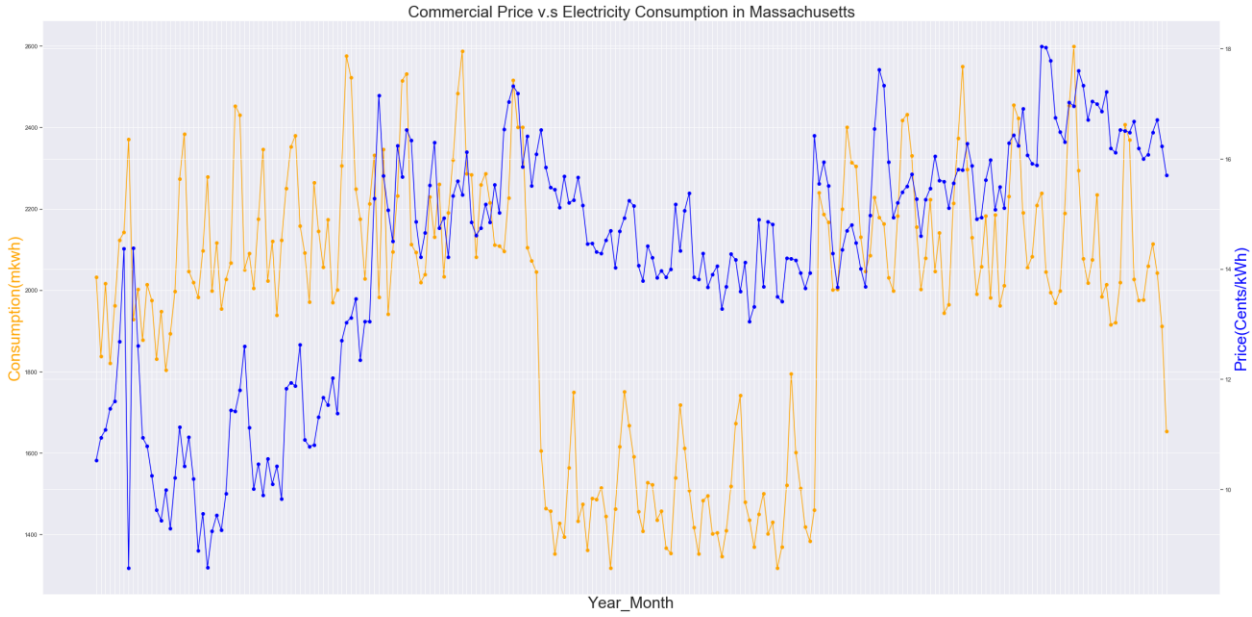
This plot shows that from 2008~2019, the time series of electricity retail price and the consumption both shows seasonal trend(the peaks are in winter). In Arizona, the shapes of the trend have the best fit upon every state. Moreover, the electricity retail price kept rising in the past decades.

North Dakota



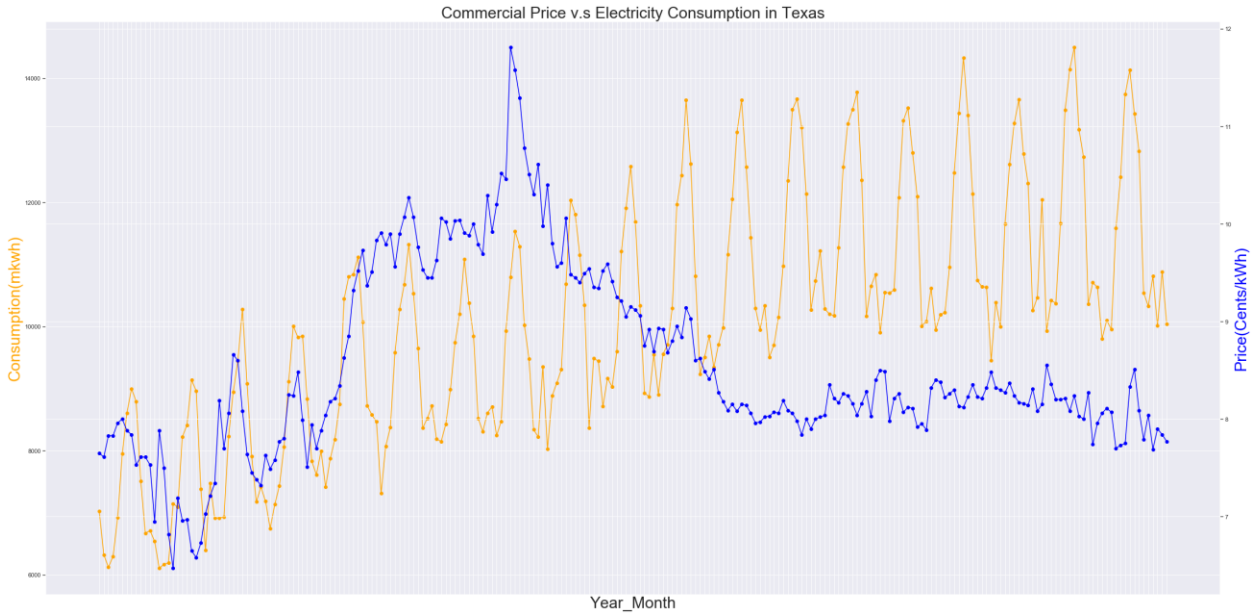
In North Dakota, the trough and the crest shows totally opposite trend, which means when the retail price is high, people who engage in commercial

activities start to use less electricity.



In Massachusetts, it shows similar phenomenon as North Dakota, when the retail price is high, people tends to use less electricity. However, the trend of Massachusetts is extreme. There's a big drop of electricity consumption from 2009~2012.

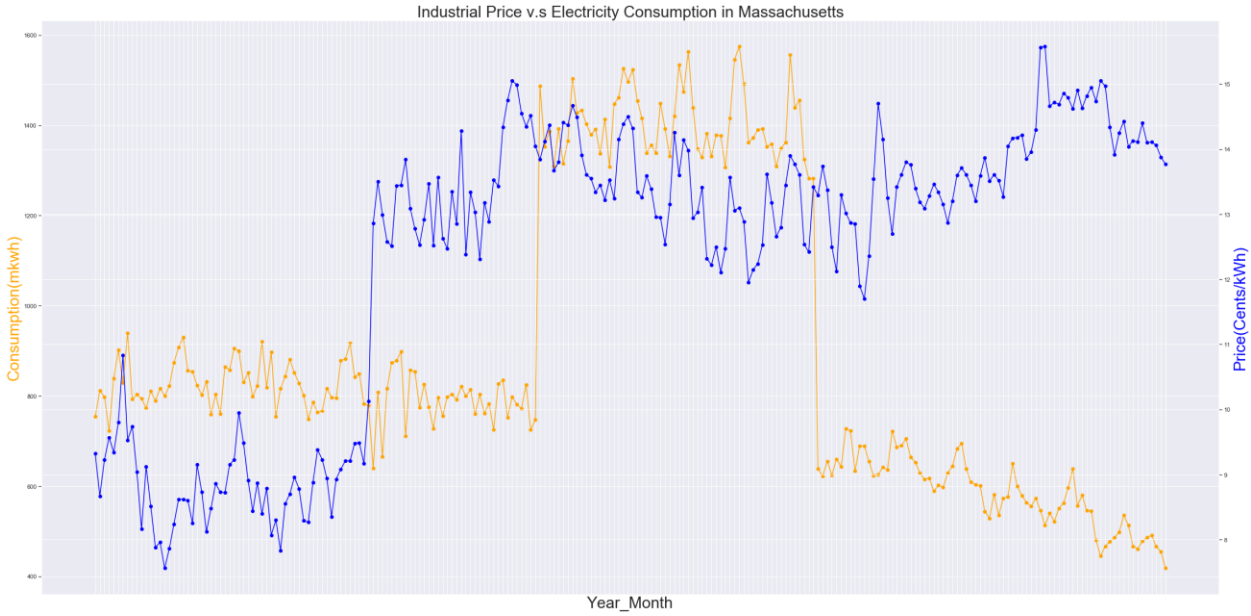
Texas



However, the retail price shows a special trend that there's a peak at 2008 and kept dropping until now.

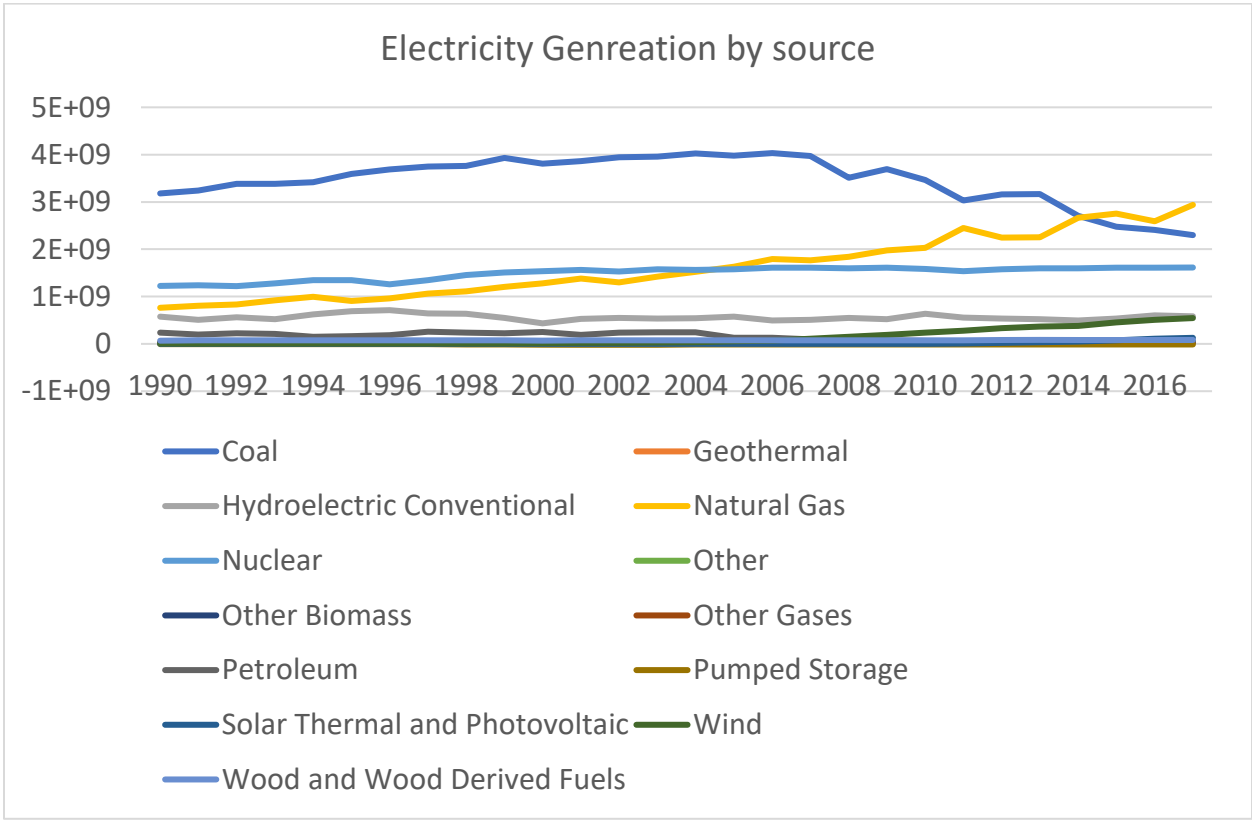
7-2. Industrial Price & Consumption

Massachusetts



Unlike the commercial electricity price and consumption trend, Massachusetts shows a extreme high peak of using industrial electricity from 2009-2012. This is an interesting phenomenon that can research deeper.

Electricity Generation by source analysis



Traditional fossil energy like: Coal, Natural Gas, Petroleum still takes up a major part of electricity generation source;

Coal has a decreasing trend along the time. Natural gas and wind are increasing along the time.

5.3.3 Model Application

5.3.3.1 Regression Analysis

I. Countrywide

- **Commercial Sector**

1. **VIF (Variance Inflation Factor)**
   We calculated the Variance Inflation Factor (VIF) for measuring collinearity among predictor variables.

```
const                    6235.708533
Commercial_Retail Price     1.227648
CLDD                       30.904525
TAVG                      471.720786
AWND                        1.112152
HTDD                      318.152483
area                        1.073584
population                  1.768300
solar-generation            1.893651
dtype: float64
```

   Among all variables, there are very high values for Monthly Average Temperature (TAVG), Heating degree days (HTDD), and Cooling Degree Days (CLDD), which indicates that they are highly correlated.

   The degree of collinearity is significantly decreased after removal of the Monthly Average Temperature (TAVG).

```
const                     24.820905
Commercial_Retail Price    1.227648
CLDD                       1.776971
AWND                       1.111356
HTDD                       1.829683
area                       1.073583
population                 1.768189
solar-generation           1.893565
dtype: float64
```

   All VIF values are smaller than 10, indicating that these variables of interests are low correlated. These variables are therefore selected for further regression analysis.

2. **Linear Regression**
   We regress the data with OLS model, and the output of regression is as follows:

OLS Regression Results

Dep. Variable:	Commercial_Usage	R-squared:	0.990
Model:	OLS	Adj. R-squared:	0.990
Method:	Least Squares	F-statistic:	4692.
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	0.00
Time:	09:08:44	Log-Likelihood:	-19911.
No. Observations:	2880	AIC:	3.994e+04
Df Residuals:	2821	BIC:	4.029e+04
Df Model:	58		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2042.5746	380.535	5.368	0.000	1296.420	2788.729
AK	-2394.8795	282.238	-8.485	0.000	-2948.294	-1841.465
AL	-1112.8947	244.481	-4.552	0.000	-1592.274	-633.516
AR	-1731.5865	301.095	-5.751	0.000	-2321.975	-1141.198
AZ	-700.1578	175.027	-4.000	0.000	-1043.352	-356.964
CA	3445.3623	813.717	4.234	0.000	1849.821	5040.904
CO	-1119.4235	216.420	-5.172	0.000	-1543.781	-695.066
CT	-1833.0572	293.002	-6.256	0.000	-2407.576	-1258.538
DE	-2177.3834	370.025	-5.884	0.000	-2902.930	-1451.837
FL	3162.9169	253.884	12.458	0.000	2665.101	3660.733
GA	395.4859	83.684	4.726	0.000	231.399	559.573
HI	-3079.7927	372.138	-8.276	0.000	-3809.484	-2350.102
IA	-1629.5684	296.511	-5.496	0.000	-2210.968	-1048.169
ID	-1863.0703	335.423	-5.554	0.000	-2520.770	-1205.371
IL	602.9546	45.866	13.146	0.000	513.021	692.888
IN	-1049.5055	193.062	-5.436	0.000	-1428.062	-670.949
KS	-1439.5274	299.524	-4.806	0.000	-2026.835	-852.220
KY	-1206.5600	258.998	-4.659	0.000	-1714.405	-698.715
LA	-987.8084	251.492	-3.928	0.000	-1480.936	-494.681
MA	-1023.4939	201.183	-5.087	0.000	-1417.974	-629.013
MD	-569.9632	217.581	-2.620	0.009	-996.598	-143.329
ME	-2166.8746	353.720	-6.126	0.000	-2860.450	-1473.299
MI	-114.6584	90.418	-1.268	0.205	-291.951	62.634
MN	-947.3627	218.715	-4.331	0.000	-1376.221	-518.505
MO	-449.7031	205.679	-2.186	0.029	-853.000	-46.406
MS	-1650.9027	301.952	-5.467	0.000	-2242.971	-1058.834
MT	-1989.7772	346.038	-5.750	0.000	-2668.291	-1311.263
NC	635.8661	89.174	7.131	0.000	461.013	810.720
ND	-1849.7877	366.212	-5.051	0.000	-2567.857	-1131.718
NE	-1741.3306	330.396	-5.270	0.000	-2389.174	-1093.488
NH	-2234.7824	357.939	-6.243	0.000	-2936.632	-1532.933
NJ	-129.9027	141.077	-0.921	0.357	-406.526	146.721
NM	-1807.7545	318.924	-5.668	0.000	-2433.102	-1182.407
NV	-1689.0413	294.505	-5.735	0.000	-2266.509	-1111.574
NY	1965.4760	214.476	9.164	0.000	1544.931	2386.021
OH	424.7019	57.429	7.395	0.000	312.094	537.310
OK	-1121.0161	271.006	-4.136	0.000	-1652.406	-589.626
OR	-1278.9643	261.447	-4.892	0.000	-1791.611	-766.318
RI	-2311.5668	368.147	-6.279	0.000	-3033.431	-1589.703
SC	-1180.8348	243.359	-4.852	0.000	-1658.015	-703.655
SD	-2009.1767	361.839	-5.553	0.000	-2718.672	-1299.682
TN	-152.4677	191.939	-0.794	0.427	-528.822	223.887
TX	6354.7232	505.274	12.577	0.000	5363.980	7345.467
UT	-1606.3914	293.775	-5.468	0.000	-2182.427	-1030.356
VA	1025.5740	140.407	7.304	0.000	750.263	1300.885
VT	-2357.8678	380.051	-6.204	0.000	-3103.075	-1612.661
WA	-495.1219	166.697	-2.970	0.003	-821.982	-168.262
WI	-931.6922	215.611	-4.321	0.000	-1354.463	-508.921
WV	-1846.2738	340.786	-5.418	0.000	-2514.488	-1178.060
WY	-2016.4992	367.516	-5.487	0.000	-2737.126	-1295.872
Commercial_Retail Price	887.4724	169.641	5.231	0.000	554.839	1220.106
CLDD	1334.3634	39.057	34.165	0.000	1257.780	1410.946
AWND	33.3686	88.006	0.379	0.705	-139.194	205.931
HTDD	370.6030	39.273	9.437	0.000	293.597	447.609
area	-13.3774	101.754	-0.131	0.895	-212.898	186.143
population	3762.6665	1190.906	3.159	0.002	1427.531	6097.802
solar-generation	56.6673	134.667	0.421	0.674	-207.388	320.723
Summer	38.0322	19.701	1.930	0.054	-0.598	76.663
Fall	80.1695	13.483	5.946	0.000	53.733	106.606
Winter	12.0360	16.557	0.727	0.467	-20.428	44.500

Omnibus:	777.712	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36033.628
Skew:	0.489	Prob(JB):	0.00
Kurtosis:	20.301	Cond. No.	3.58e+15

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.07e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- R-Squared, Adj R-Squared
  The value of R-Squared is 0.990, adjusted R-Squared is also 0.990. This indicates that after accounting for predictors that are not significant in a regression model, 99.0% of commercial electricity consumption can be explained by the independent variables.

- MAE, MSE, RMSE

```
Performance Evaluation
Mean Absolute Error: 139.59305809080473
Mean Squared Error: 60937.57972502659
Root Mean Squared Error: 246.85538220793686
```

- Feature coefficients and Significance

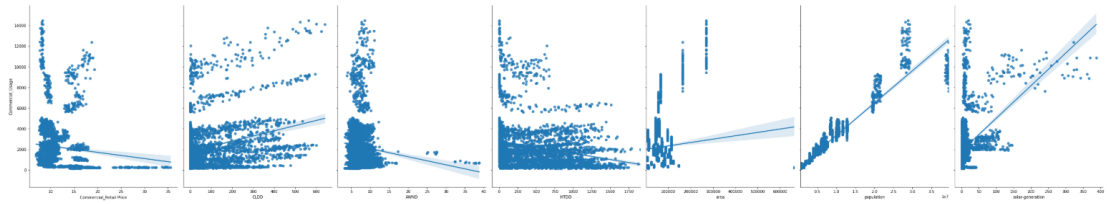
Feature	**Feature coefficients and Significance**
Commercial Retail Price	Statistically significant. A near zero P-Value (P=0.000) indicates that there is sufficient evidence to suggest that retail price (Cents/kWh) has influence. On average, an increase of 1 cent per kWh will increase national commercial consumption by 887 million kilowatt-hours.
CLDD	Statistically significant. A near zero P-Value (P=0.000) indicates that there is sufficient evidence to suggest that Cooling Degree Days (CLDD) has influence. On average, an increase of the reduction of 1 ºF will increase national commercial consumption by 1334 million kilowatt-hours.
AWND	Not statistically significant. A high P-Value (P=0.705) indicates that there is no sufficient evidence to suggest that Average wind speed (AWND) has influence.
HTDD	Statistically significant. A near zero P-Value (P=0.000) indicates that there is sufficient evidence to suggest that Heating degree days (HTDD) has influence.

	On average, an increase of 1 ºF will increase commercial consumption by 371 million kilowatt-hours.
Area	Not statistically significant. A relatively high P-Value (P=0.895) indicates that there is no sufficient evidence to suggest that state area (area) has influence.
Population	Statistically significant. A near zero P-Value (P=0.002) indicates that there is sufficient evidence to suggest that state population has influence. On average, an increase of 1 person in the state will increase commercial consumption by 3763 million kilowatt-hours.
Solar-Generation	Not statistically significant. A high P-Value (P=0.674) indicates that there is no sufficient evidence to suggest that state solar-generation has influence.
Season: Summer, Season: Fall, Season: Winter	Not statistically significant. Some high P-values ($P \geq 0.05$) for these three dummy variables indicating that seasonal factors not surely have influence on industrial electricity consumption.
CA	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that CA's monthly demand for commercial consumption demand varies from PA. On average, CA's commercial sector consumes 3445 million kilowatt-hours more than PA per month.
FL	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that FL's monthly

	demand for commercial consumption demand varies from PA. On average, FL's commercial sector consumes 3163 million kilowatt-hours more than PA per month.
MN	Statistically significant. A relatively low P-Value (P=0.000) suggests that there is sufficient evidence to conclude that MN's monthly demand for commercial consumption demand varies from PA. On average, MN's commercial sector consumes 947 million kilowatt-hours less than PA per month.
MO	Statistically significant. A relatively low P-Value (P=0.029) suggests that there is sufficient evidence to conclude that MO's monthly demand for commercial consumption demand varies from PA. On average, MO's commercial sector consumes 450 million kilowatt-hours less than PA per month.
NC	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that NC's monthly demand for commercial consumption demand varies from PA. On average, NC's commercial sector consumes 636 million kilowatt-hours more than PA per month.
NY	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that NY's monthly demand for commercial consumption demand varies from PA.

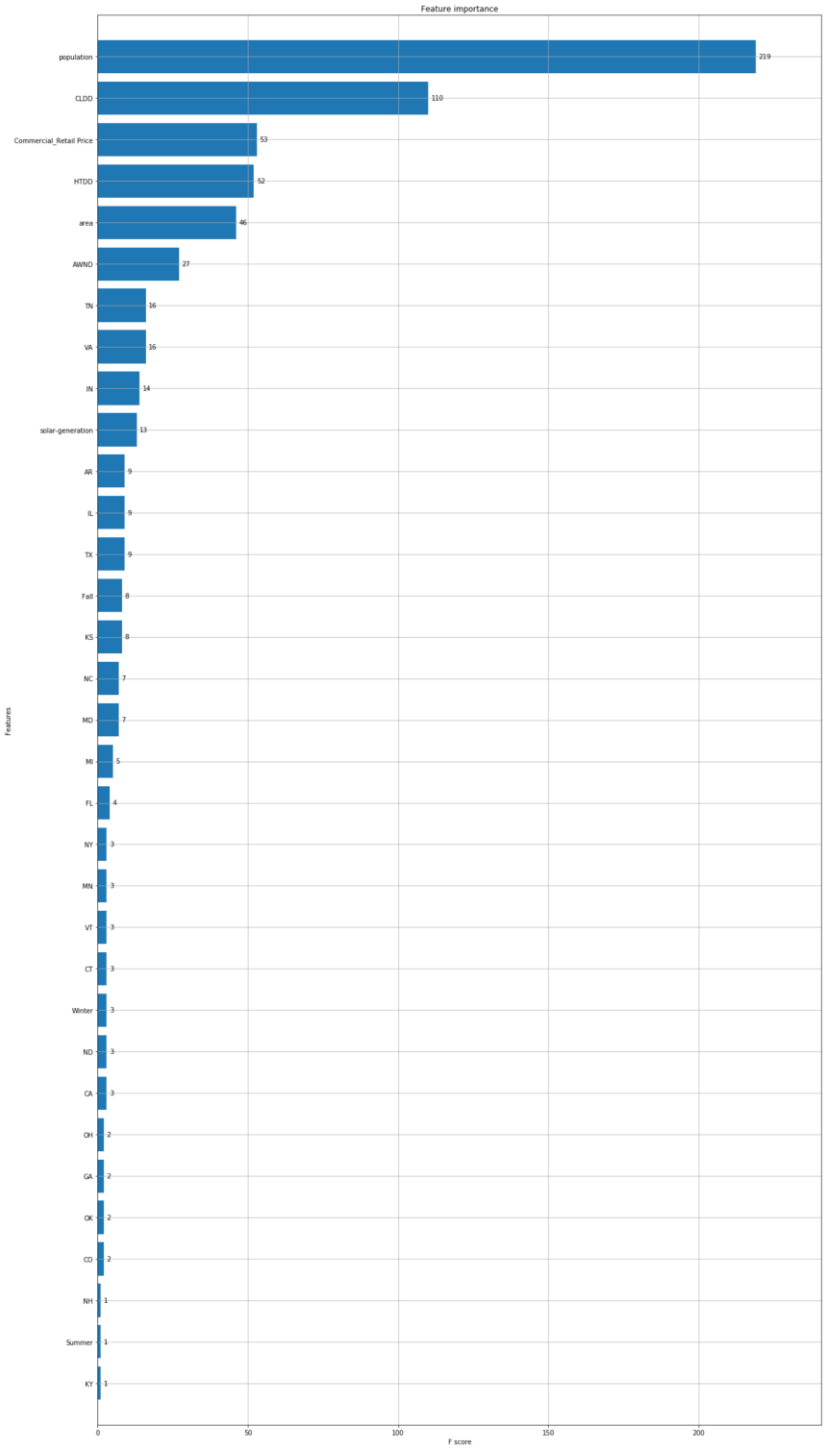
	On average, NY's commercial sector consumes 1965 million kilowatt-hours more than PA per month.
PA	Holding other things constant, PA's monthly commercial consumption is 2043 million kilowatt-hours.
TX	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that TX's monthly demand for commercial consumption demand varies from PA. On average, TX's commercial sector consumes 6355 million kilowatt-hours more than PA per month.
WA	Statistically significant. A near zero P-Value (P=0.003) suggests that there is sufficient evidence to conclude that WA's monthly demand for commercial consumption demand varies from PA. On average, WA's commercial sector consumes 495 million kilowatt-hours less than PA per month.
WY	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that WY's monthly demand for commercial consumption demand varies from PA. On average, WY's commercial sector consumes 2016 million kilowatt-hours less than PA per month.

- Feature Correlation
  Population and solar-generation two independent variables appear to have linear correlation with dependent variable monthly national commercial electricity consumption.

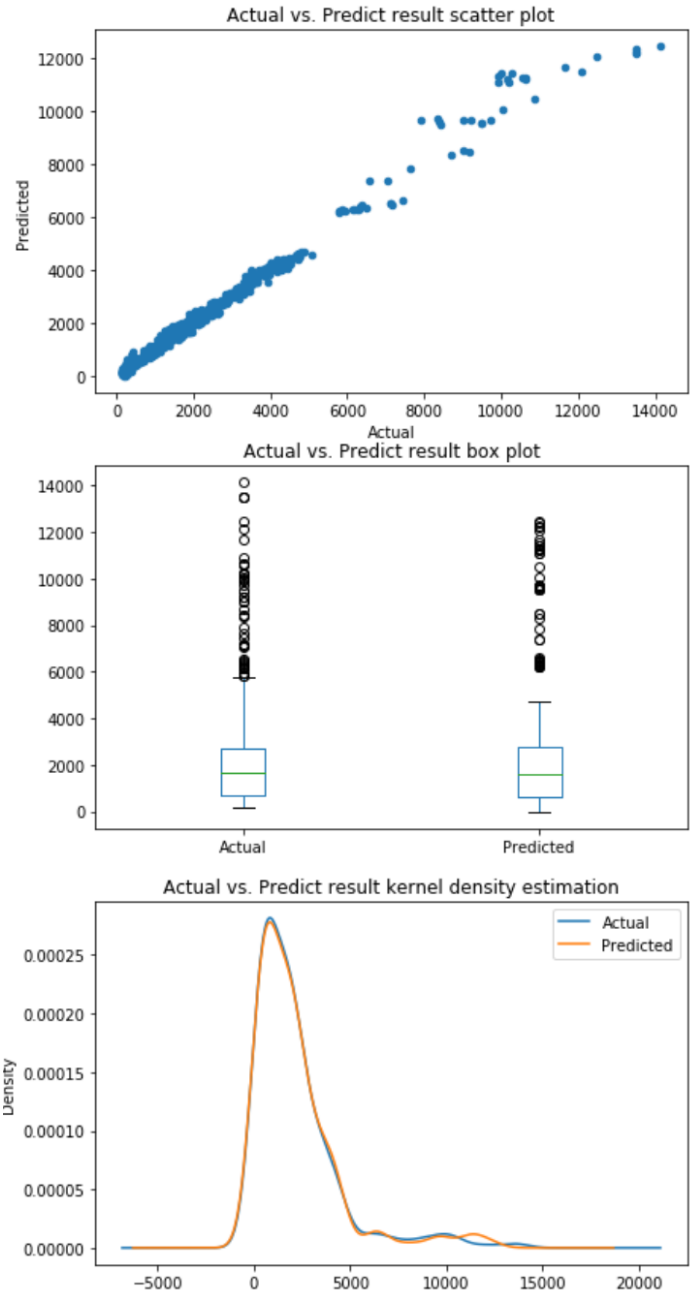


- Feature Importance
  In the commercial sector, population is the most important factor for electricity demand. This is quite intuitive, commercial electricity consumption is strongly related to number of people. When the number of people increases, commercial activities should grow accordingly. Hence, the demand for electricity in the commercial sector will also increase.

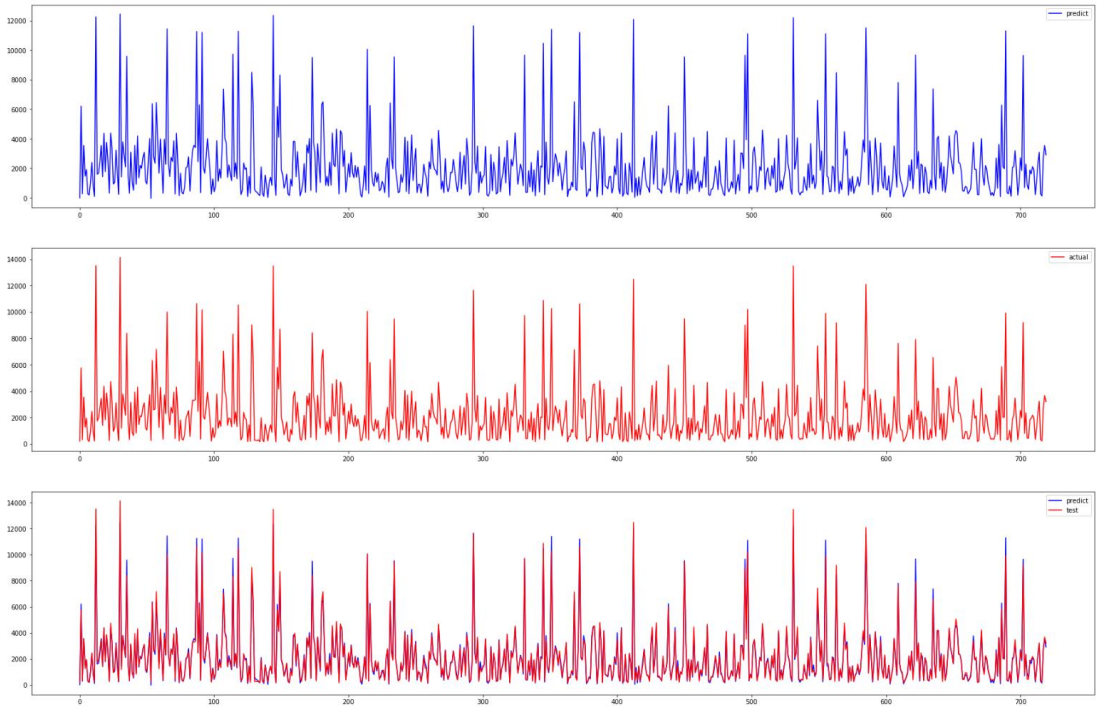
  Cooling Degree Days (CLDD) is the second most important factor for electricity demand. On average, an increase in the 1 ºF reduction would increase commercial consumption by 1334 million kilowatt-hours according to the outcome of linear regression.



- Actual vs. Predict Visualization
  After visualizing the result in boxplot, scatter plot, and estimation of kernel density, we can conclude that the predicted value is very similar to the actual value, on average.



- Actual vs Predict Plot



In the third graph, the red line represents the observed value, and the blue line represents the predicted value. Through this graph, we can tell that the difference between the actual value and the predicted value is relatively low.

- **Industrial Sector**

1. **VIF (Variance Inflation Factor)**
   We calculated the Variance Inflation Factor (VIF) for measuring collinearity among predictor variables.

```
const              6228.644523
Cents/kWh             1.129253
CLDD                 30.902087
TAVG                471.704632
AWND                  1.119652
HTDD                318.153807
area                  1.068077
Population            1.791753
solar-generation      1.781024
dtype: float64
```

   Among all variables, there are very high values for Monthly Average Temperature (TAVG), Heating degree days (HTDD), and Cooling Degree Days (CLDD), which indicates that they are highly correlated.

The degree of collinearity is significantly decreased after removal of the Monthly Average Temperature (TAVG).

```
const              19.440557
Cents/kWh           1.129237
CLDD                1.787083
AWND                1.118834
HTDD                1.823300
area                1.068077
Population          1.791727
solar-generation    1.781023
dtype: float64
```

All VIF values are smaller than 10, indicating that these variables of interests are low correlated. These variables are therefore selected for further regression analysis.

2. Linear Regression

We regress the data with OLS model, and the output of regression is as follows:

OLS Regression Results

Dep. Variable:	industrial_usage	R-squared:	0.992
Model:	OLS	Adj. R-squared:	0.991
Method:	Least Squares	F-statistic:	5698.
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	0.00
Time:	13:53:17	Log-Likelihood:	-18465.
No. Observations:	2880	AIC:	3.705e+04
Df Residuals:	2821	BIC:	3.740e+04
Df Model:	58		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3100.1840	250.073	12.397	0.000	2609.840	3590.528
Cents/kWh	526.4793	105.505	4.990	0.000	319.605	733.354
CLDD	308.6080	24.121	12.794	0.000	261.311	355.905
AWND	-27.2517	54.300	-0.502	0.616	-133.723	79.220
HTDD	34.2815	23.627	1.451	0.147	-12.046	80.608
area	-1360.2926	65.697	-20.706	0.000	-1489.111	-1231.474
AK	-1892.2170	184.597	-10.251	0.000	-2254.175	-1530.259
AL	-667.8429	160.363	-4.165	0.000	-982.283	-353.403
AR	-1891.2131	197.450	-9.578	0.000	-2278.373	-1504.053
AZ	-2286.3787	112.023	-20.410	0.000	-2506.034	-2066.723
CA	-1735.4025	533.442	-3.253	0.001	-2781.379	-689.426
CO	-2062.6704	140.420	-14.689	0.000	-2338.006	-1787.334
CT	-3260.3909	190.565	-17.109	0.000	-3634.053	-2886.729
DE	-3081.0815	242.168	-12.723	0.000	-3555.926	-2606.237
FL	-3308.4437	166.852	-19.829	0.000	-3635.609	-2981.279
GA	-1168.0733	54.541	-21.416	0.000	-1275.018	-1061.129
HI	-3350.7982	238.850	-14.029	0.000	-3819.136	-2882.460
IA	-1402.5199	194.125	-7.225	0.000	-1783.162	-1021.878
ID	-2387.7491	219.809	-10.863	0.000	-2818.752	-1956.746
IL	-390.1318	28.280	-13.795	0.000	-445.584	-334.680
IN	189.2638	126.167	1.500	0.134	-58.124	436.652
KS	-2297.4209	195.973	-11.723	0.000	-2681.687	-1913.155
KY	-964.4545	169.798	-5.680	0.000	-1297.396	-631.513
LA	-466.1992	164.960	-2.826	0.005	-789.654	-142.744
MA	-3171.0022	126.885	-24.991	0.000	-3419.799	-2922.205
MD	-3309.5125	141.376	-23.409	0.000	-3586.722	-3032.303
ME	-2970.7674	231.357	-12.841	0.000	-3424.414	-2517.121
MI	-1138.9121	58.031	-19.626	0.000	-1252.699	-1025.126
MN	-1574.5822	142.998	-11.011	0.000	-1854.973	-1294.191
MO	-2302.1743	134.247	-17.149	0.000	-2565.406	-2038.942
MS	-1976.6511	198.132	-9.976	0.000	-2365.150	-1588.152
MT	-2520.0985	227.469	-11.079	0.000	-2966.120	-2074.077
NC	-1552.9802	56.719	-27.380	0.000	-1664.195	-1441.766
ND	-2408.0652	239.713	-10.046	0.000	-2878.095	-1938.035
NE	-2259.4995	216.130	-10.454	0.000	-2683.289	-1835.710
NH	-3173.6783	233.861	-13.571	0.000	-3632.234	-2715.122
NJ	-3269.4149	87.432	-37.394	0.000	-3440.853	-3097.977
NM	-2388.5827	208.774	-11.441	0.000	-2797.947	-1979.218
NV	-2051.4026	192.510	-10.656	0.000	-2428.878	-1673.927
NY	-2992.7874	139.427	-21.465	0.000	-3266.177	-2719.398
OH	223.0685	36.071	6.184	0.000	152.339	293.798
OK	-1774.7735	177.549	-9.996	0.000	-2122.912	-1426.635
OR	-2151.1492	170.672	-12.604	0.000	-2485.803	-1816.495
RI	-3301.0386	240.606	-13.720	0.000	-3772.820	-2829.258
SC	-1185.9322	159.598	-7.431	0.000	-1498.873	-872.991
SD	-2836.1831	236.873	-11.973	0.000	-3300.645	-2371.721
TN	-1712.4625	125.336	-13.663	0.000	-1958.222	-1466.703
TX	4983.4321	331.596	15.029	0.000	4333.236	5633.628
UT	-2414.3110	192.479	-12.543	0.000	-2791.725	-2036.896
VA	-2261.7878	91.353	-24.759	0.000	-2440.912	-2082.663
VT	-3124.8582	247.913	-12.605	0.000	-3610.966	-2638.750
WA	-1358.6120	109.410	-12.418	0.000	-1573.143	-1144.081
WI	-1449.0616	140.720	-10.297	0.000	-1724.986	-1173.137
WV	-2115.0820	223.210	-9.476	0.000	-2552.753	-1677.411
WY	-2128.1976	240.813	-8.838	0.000	-2600.384	-1656.011
Population	2864.2893	780.426	3.670	0.000	1334.026	4394.553
solar-generation	207.1094	93.420	2.217	0.027	23.932	390.287
Season_Fall	9.7680	8.084	1.208	0.227	-6.083	25.619
Season_Summer	23.9398	11.968	2.000	0.046	0.473	47.407
Season_Winter	-47.3965	10.118	-4.684	0.000	-67.237	-27.556

Omnibus:	586.391	Durbin-Watson:	2.076
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12784.012
Skew:	0.382	Prob(JB):	0.00
Kurtosis:	13.293	Cond. No.	4.72e+16

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.78e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- R-Squared, Adj R-Squared
  The value of R-Squared is 0.992, adjusted R-Squared is 0.991. This indicates that after accounting for predictors that are not significant in a regression model, 99.1% of industrial electricity consumption can be explained by the independent variables.

- MAE, MSE, RMSE

```
Mean Absolute Error: 96.86851399739584
Mean Squared Error: 29161.301314634748
Root Mean Squared Error: 170.7668039012113
```

- Feature coefficients and Significance

Feature	**Feature coefficients and Significance**
Cents/kWh	Statistically significant. A near zero P-Value (P=0.000) indicates that there is sufficient evidence to suggest that retail price (Cents/kWh) has influence. On average, an increase of 1 cent per kWh will increase industrial consumption by 526 million kilowatt-hours.
CLDD	Statistically significant. A near zero P-Value (P=0.000) indicates that there is sufficient evidence to suggest that Cooling Degree Days (CLDD) has influence. On average, an increase of the reduction of 1 ºF will increase industrial consumption by 309 million kilowatt-hours.
AWND	Not statistically significant. A high P-Value (P=0.616) indicates that there is no sufficient evidence to suggest that Average wind speed (AWND) has influence.
HTDD	Not statistically significant. A high P-Value (P=1.451) indicates that there is no sufficient evidence to suggest that Heating degree days (HTDD) has influence.
Area	Statistically significant. A near zero P-Value (P=0.000) indicates that there is sufficient

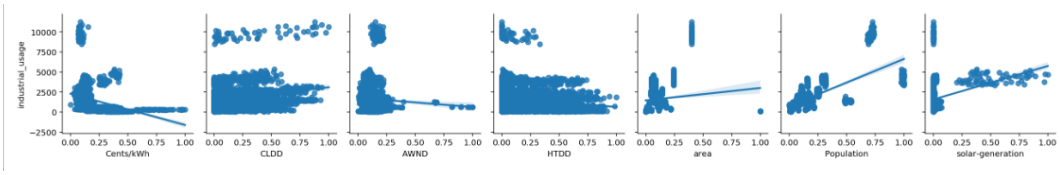
	evidence to suggest that state area (area) has influence. On average, an increase of 1 square mile will decrease industrial consumption by 1360 million kilowatt-hours.
Population	Statistically significant. A near zero P-Value (P=0.000) indicates that there is sufficient evidence to suggest that state population has influence. On average, an increase of 1 person in the state will increase industrial consumption by 2864 million kilowatt-hours.
Solar-Generation	Statistically significant. A relatively low P-Value (P=0.027) indicates that there is sufficient evidence to suggest that state solar-generation has influence. On average, an increase of 1 million kilowatt-hours of electricity generated by distributed solar-PV, will increase industrial consumption by 207 million kilowatt-hours.
Season: Summer, Season: Fall, Season: Winter	Not statistically significant. Some high P-values ($P \geq 0.05$) for these three dummy variables indicating that seasonal factors not surely have influence on industrial electricity consumption.
CA	Statistically significant. A near zero P-Value (P=0.001) suggests that there is sufficient evidence to conclude that CA's monthly demand for industrial consumption demand varies from PA. On average, CA's industrial sector consumes 1735 million kilowatt-hours less than PA per month.

FL	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that FL's monthly demand for industrial consumption demand varies from PA. On average, FL's industrial sector consumes 3308 million kilowatt-hours less than PA per month.
MN	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that MN's monthly demand for industrial consumption demand varies from PA. On average, MN's industrial sector consumes 1575 million kilowatt-hours less than PA per month.
MO	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that MO's monthly demand for industrial consumption demand varies from PA. On average, MO's industrial sector consumes 2302 million kilowatt-hours less than PA per month.
NC	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that NC's monthly demand for industrial consumption demand varies from PA. On average, NC's industrial sector consumes 1553 million kilowatt-hours less than PA per month.
NY	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that NY's monthly

	demand for industrial consumption demand varies from PA. On average, NY's industrial sector consumes 2993 million kilowatt-hours less than PA per month.
PA	Holding other things constant, PA's monthly industrial consumption is 3100 million kilowatt-hours.
TX	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that TX's monthly demand for industrial consumption demand varies from PA. On average, TX's industrial sector consumes 4983 million kilowatt-hours more than PA per month.
WA	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that WA's monthly demand for industrial consumption demand varies from PA. On average, WA's industrial sector consumes 1359 million kilowatt-hours less than PA per month.
WY	Statistically significant. A near zero P-Value (P=0.000) suggests that there is sufficient evidence to conclude that WY's monthly demand for industrial consumption demand varies from PA. On average, WY's industrial sector consumes 2128 million kilowatt-hours less than PA per month.

- Feature Correlation

Population and solar-generation two independent variables appear to have linear correlation with dependent variable monthly national industrial electricity usage.



- Feature Importance
  In the industrial sector, price is the most important factor for electricity demand. This is quite intuitive, as industrial electricity consumption is sensitive to electricity prices, as this will be reflected in the cost of production.

  Population is the second most important factor for industrial electricity demand. On average, an increase in the 1 person in the state will increase industrial consumption by 2864 million kilowatt-hours according to the outcome of linear regression.

- **Residential Sector**

1. VIF
   We calculated the VIF of all numeric independent variables in order to check for multicollinearity. The calculated VIF is as follows:

```
const               6229.832646
Cents/kWh             1.081414
CLDD                 30.900533
TAVG                471.845143
AWND                  1.113365
HTDD                318.196635
area                  1.057249
Population            1.722443
solar-generation      1.746774
dtype: float64

-------------After removing TAVG--------------

const               26.076614
Cents/kWh            1.081123
CLDD                 1.766007
AWND                 1.112518
HTDD                 1.832883
area                 1.057248
Population           1.722393
solar-generation     1.746764
dtype: float64
```

We could observe that variable TAVG, HTDD and area have a very large VIF, indicating that there might be multicollinearity among the variables. We try to remove TAVG and area to eliminate multicollinearity, and the VIF after removal are smaller than 10, indicating that there is no multicollinearity among the independent variables now.

2. Residential Consumption

We regress the data with OLS model, and the output of regression is as follows:

OLS Regression Results

Dep. Variable:	residential_usage	R-squared:	0.957
Model:	OLS	Adj. R-squared:	0.956
Method:	Least Squares	F-statistic:	1084.
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	0.00
Time:	04:44:59	Log-Likelihood:	-22087.
No. Observations:	2880	AIC:	4.429e+04
Df Residuals:	2821	BIC:	4.464e+04
Df Model:	58		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1413.6911	839.186	-1.685	0.092	-3059.171	231.788
Cents/kWh	-657.7090	323.432	-2.034	0.042	-1291.896	-23.522
CLDD	2541.2510	83.154	30.561	0.000	2378.203	2704.299
AWND	-22.0242	192.205	-0.115	0.909	-398.901	354.853
HTDD	1271.7384	84.438	15.061	0.000	1106.171	1437.306
area	4.857e-05	0.001	0.078	0.938	-0.001	0.001
AK	980.7598	453.870	2.161	0.031	90.809	1870.710
AL	1315.0883	529.999	2.481	0.013	275.864	2354.312
AR	956.0020	652.234	1.466	0.143	-322.902	2234.906
AZ	582.8490	361.799	1.611	0.107	-126.568	1292.266
CA	-8808.1074	1783.517	-4.939	0.000	-1.23e+04	-5310.978
CO	157.3879	453.332	0.347	0.728	-731.508	1046.283
CT	703.0852	650.859	1.080	0.280	-573.122	1979.292
DE	863.5527	816.306	1.058	0.290	-737.065	2464.170
FL	1521.9270	558.378	2.726	0.006	427.057	2616.797
GA	1071.8495	178.780	5.995	0.000	721.297	1422.402
HI	583.4879	814.367	0.716	0.474	-1013.327	2180.303
IA	784.1192	640.765	1.224	0.221	-472.295	2040.534
ID	966.7399	717.520	1.347	0.178	-440.177	2373.657
IL	-774.0799	100.158	-7.729	0.000	-970.470	-577.690
IN	801.2680	422.023	1.899	0.058	-26.237	1628.773
KS	719.8634	640.573	1.124	0.261	-536.176	1975.903
KY	1150.8149	564.166	2.040	0.041	44.594	2257.035
LA	1095.3864	547.000	2.003	0.045	22.827	2167.946
MA	-123.5045	436.908	-0.283	0.777	-980.195	733.186

MD	595.0331	482.019	1.234	0.217	-350.112	1540.178
ME	929.6252	772.971	1.203	0.229	-586.021	2445.272
MI	-363.4092	183.343	-1.982	0.048	-722.909	-3.910
MN	420.1980	464.470	0.905	0.366	-490.538	1330.934
MO	1095.9010	440.464	2.488	0.013	232.237	1959.565
MS	957.0242	655.945	1.459	0.145	-329.157	2243.205
MT	945.5208	724.143	1.306	0.192	-474.383	2365.424
NC	1300.8326	192.154	6.770	0.000	924.055	1677.610
ND	1041.7939	788.038	1.322	0.186	-503.394	2586.982
NE	951.5213	708.333	1.343	0.179	-437.382	2340.424
NH	954.3085	790.662	1.207	0.228	-596.026	2504.643
NJ	-431.2677	304.141	-1.418	0.156	-1027.630	165.094
NM	639.6646	671.811	0.952	0.341	-677.626	1956.955
NV	799.2116	621.698	1.286	0.199	-419.816	2018.240
NY	-3013.5640	460.792	-6.540	0.000	-3917.087	-2110.041
OH	351.5420	124.246	2.829	0.005	107.920	595.164
OK	937.5869	581.878	1.611	0.107	-203.362	2078.536
OR	941.5604	551.511	1.707	0.088	-139.845	2022.965
RI	973.7408	814.310	1.196	0.232	-622.962	2570.444
SC	1172.2039	533.908	2.196	0.028	125.313	2219.094
SD	951.1380	776.873	1.224	0.221	-572.158	2474.435
TN	1435.7649	417.063	3.443	0.001	617.986	2253.544
TX	624.2687	1153.119	0.541	0.588	-1636.773	2885.311
UT	405.9682	628.300	0.646	0.518	-826.006	1637.942
VA	1167.8756	305.979	3.817	0.000	567.911	1767.840
VT	1036.3742	836.880	1.238	0.216	-604.584	2677.333
WA	790.7426	358.454	2.206	0.027	87.884	1493.601
WI	349.8817	462.690	0.756	0.450	-557.364	1257.127
WV	1136.3176	745.630	1.524	0.128	-325.717	2598.352
WY	985.8808	783.576	1.258	0.208	-550.558	2522.320
Population	1.672e+04	2586.140	6.464	0.000	1.16e+04	2.18e+04
solar-generation	843.8600	313.407	2.693	0.007	229.330	1458.390
Season_Fall	111.7177	28.647	3.900	0.000	55.547	167.888
Season_Summer	271.3904	41.495	6.540	0.000	190.026	352.755
Season_Winter	330.1262	35.398	9.326	0.000	260.718	399.535

Omnibus:	965.741	Durbin-Watson:	1.917
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22983.164
Skew:	1.035	Prob(JB):	0.00
Kurtosis:	16.683	Cond. No.	1.24e+19

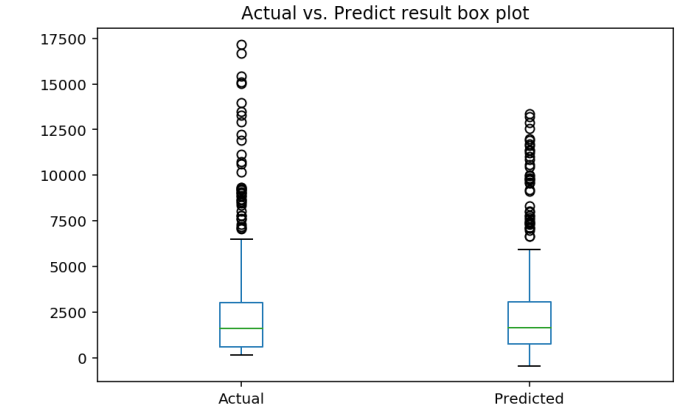
- R-squared, Adjusted R-squared
  The value of R-squared is 0.957, adjusted R-squared is 0.956. This indicates that after adjusting for the number of predictors, 95.6% of the residential energy consumption of nationwide could be explained by the independent variables.

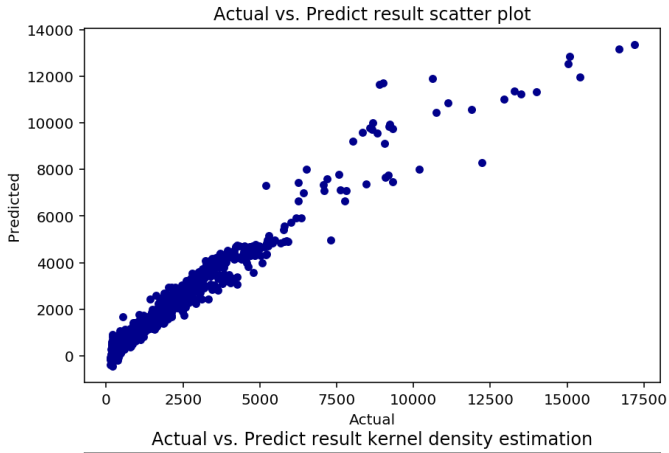
- Feature Coefficients and Significance
  Among the independent variables, we could see clearly that there are some significant features that have an impact on the commercial energy consumption. “CLDD” and “HTDD” are two statistically significant weather indicators, indicating that when indoors cooling days increase by 1, average monthly commercial consumption increase by 2541 kilo-watthours, when heating days increase by 1, average monthly residential consumption increase by 1271 kilo-watthours. Retail Price is another significant indicator, showing that when average retail price increase by 1cents per kilo-watthours, the overall residential consumption decrease by 657 kilo-watthours. Season dummy variable are also statistically significant. The residential electricity consumption in winter, summer and fall are expected to be 330 kilo-watthours, 271 kilo-watthours and 111 kilo-watthours higher than in Spring. The majority of state indicators are not significant under 0.05 significant level, while states New York, California and

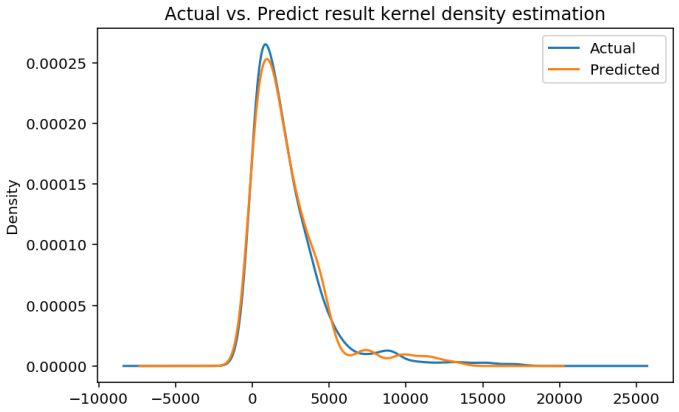
Virginia are strongly significant, showing residential usage in these states are statistically differentiate with Pennsylvania, the baseline state.

- MAE, MSE, RMSE
  Mean Absolute Error: 354.40658000469386
  Mean Squared Error: 314057.22003811045
  Root Mean Squared Error: 560.4080834874801

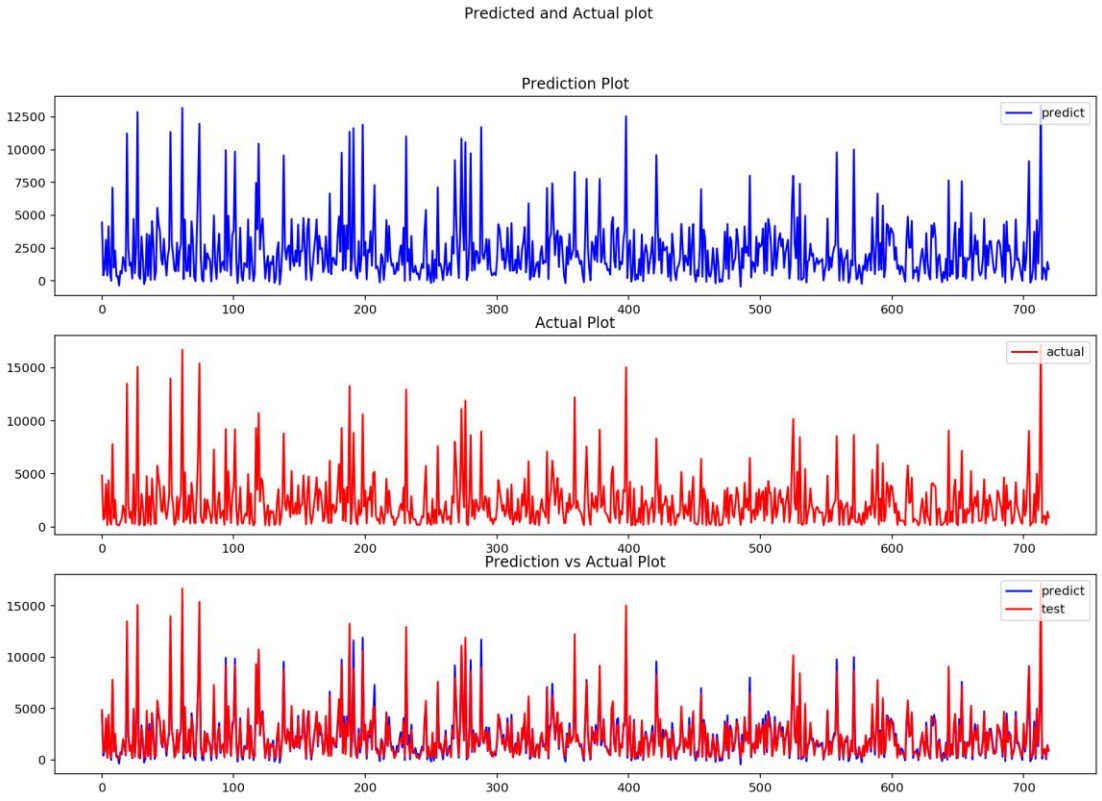
- Actual vs. Prediction Visualization







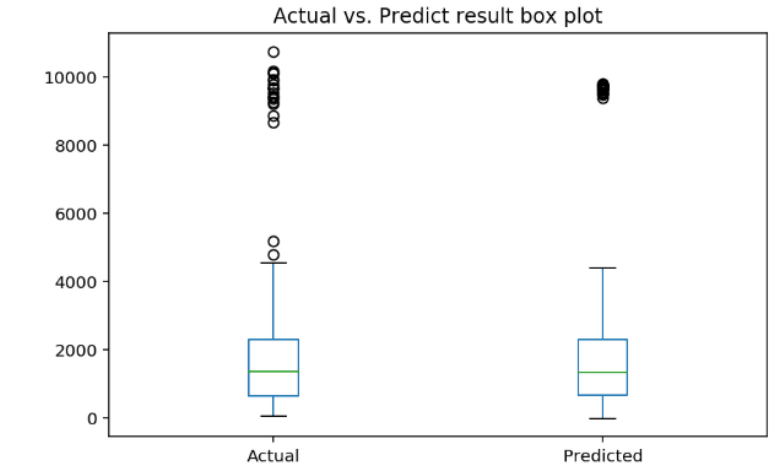
- Actual vs. Prediction plot
  From the box plot we see predicted value have lower variance than the true value. From scatter plot, we found the predicted value generally are lower than the real one. The Distribution of both predicted and true value are of the same shape.

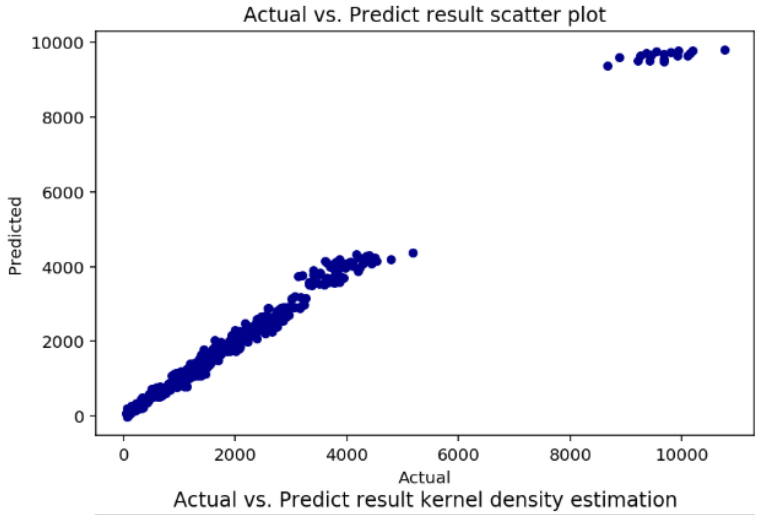


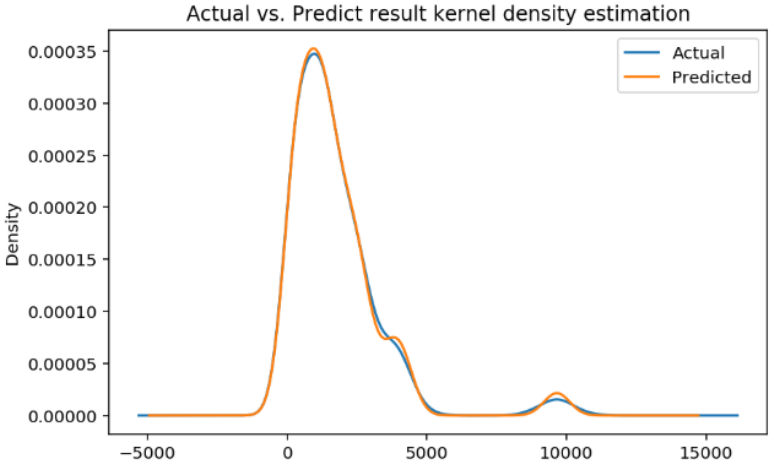
From the visualization graphs we could see that the model is doing a relatively good job in predicting residential energy consumption. The errors are sparsely distributed at the whole prediction range. When predictions are below 100 or above 600,


Feature importance
Cents/kWh 906
Population 778
AWND 525
CLDD 469
HTDD 302
area 96
solar-generation 72
Season_Fall 37
Season_Winter 37
ID 27
IN 19
GA 15
MO 14
WI 13
NV 13
UT 13
IL 12
MN 12
VA 12
NE 12
IA 12
KY 11
OH 11
TN 11
TX 11
MS 10
NY 10
MI 9
AZ 9
NC 9
SC 8
AR 8
SD 8
Season_Summer 8
LA 8
ND 8
WY 8
AL 7
FL 7
WA 6
WV 5
MD 5
CO 5
HI 4
OR 3
ME 3
KS 3
NM 3
NJ 2
MA 2
MT 2
OK 2
CT 2
NH 1
Features
0
200
400
600
800
F score

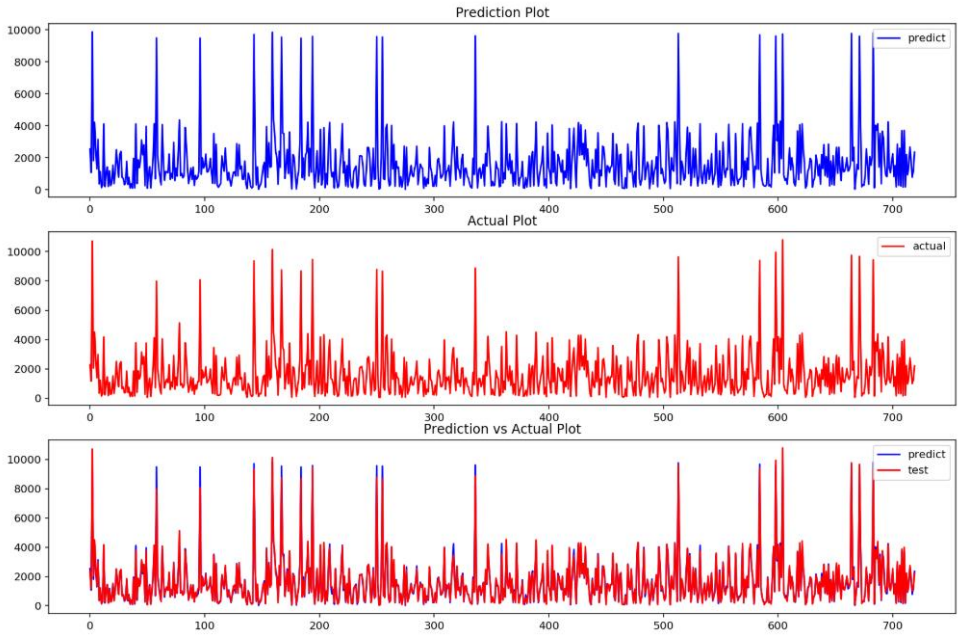
- Actual vs. Predict Visualization
  After visualizing the result in boxplot, scatter plot, and estimation of kernel density, we can conclude that the predicted value is very similar to the actual value, on average.







- Actual vs Predict Plot
  In the third graph, the red line represents the observed value, and the blue line represents the predicted value. Through this graph, we can tell that the difference between the actual value and the predicted value is relatively low.



II. State-Wise

1. Pennsylvania

1.1 Commercial Sector for Pennsylvania

1.1.1 VIF

```
Retail Price(Cents/kWh)    2.876084e+00
TAVG                       4.302090e+02
area                       7.114711e+06
CLDD                       1.399268e+01
AWND                       3.787500e+00
HTDD                       3.235793e+02
population                 2.416940e+00
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove TAVG and area to reduce the multicollinearity from the dataset.

```
const                      6.440960e+06
Retail Price(Cents/kWh)    2.498478e+00
CLDD                       2.942709e+00
AWND                       3.787218e+00
HTDD                       3.804797e+00
population                 2.230025e+00
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

1.1.2 Regression on the Commercial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	mkwh	R-squared:	0.801
Model:	OLS	Adj. R-squared:	0.762
Method:	Least Squares	F-statistic:	20.97
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	1.23e-13
Time:	12:29:01	Log-Likelihood:	-345.03
No. Observations:	57	AIC:	710.1
Df Residuals:	47	BIC:	730.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.236e+05	5.89e+04	2.099	0.041	5157.896	2.42e+05
Summer	-40.8751	74.807	-0.546	0.587	-191.368	109.618
Fall	25.7027	71.304	0.360	0.720	-117.743	169.148
Winter	-7.7830	74.431	-0.105	0.917	-157.518	141.952
Retail Price(Cents/kWh)	-14.6940	632.272	-0.023	0.982	-1286.661	1257.273
CLDD	1037.3410	150.319	6.901	0.000	734.939	1339.743
AWND	-606.6351	279.928	-2.167	0.035	-1169.778	-43.493
HTDD	862.7721	169.045	5.104	0.000	522.698	1202.846
population	-1.203e+05	5.87e+04	-2.048	0.046	-2.38e+05	-2144.726
solar-generation	144.5518	277.487	0.521	0.605	-413.680	702.783

Omnibus:	0.729	Durbin-Watson:	1.577
Prob(Omnibus):	0.694	Jarque-Bera (JB):	0.201
Skew:	-0.036	Prob(JB):	0.904
Kurtosis:	3.282	Cond. No.	1.11e+04

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.801, and the adjusted R-squared is 0.762, there is not much difference in them, so we are safe to say that our

input variables are all valuable to the model, and they can explain 80.1% of the overall result.

Feature coefficients and Significance

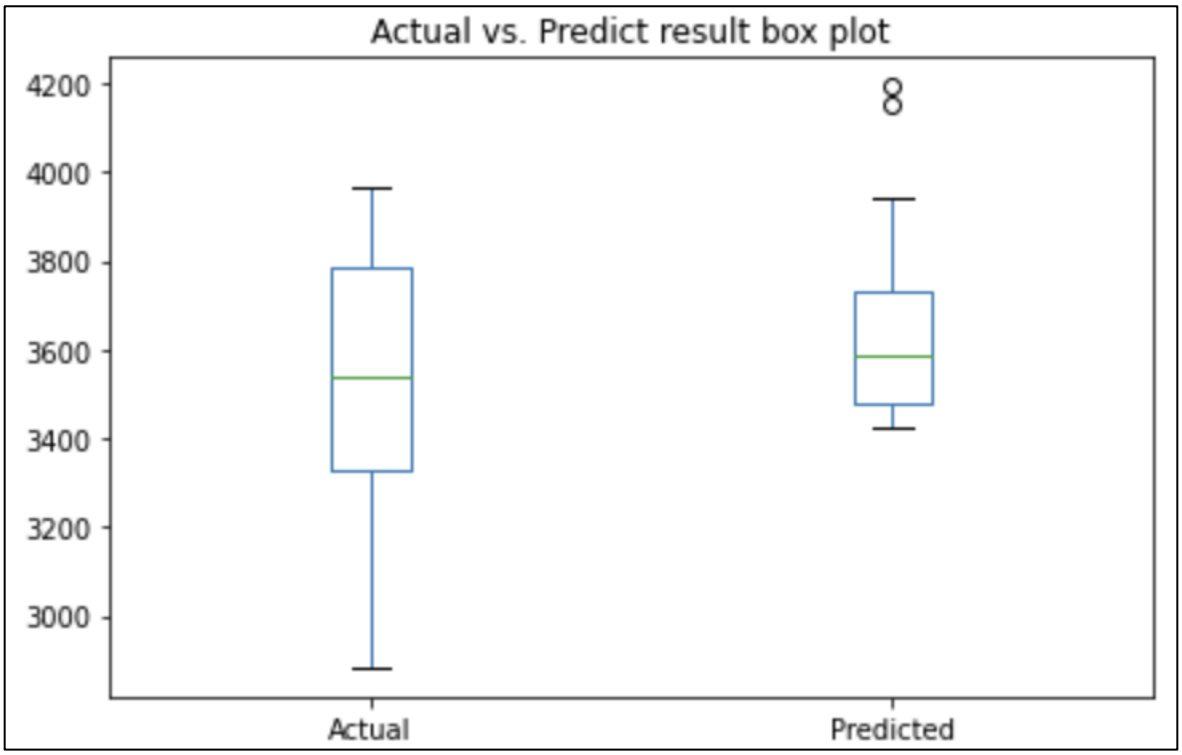
From the above table, we can identify some features that have significant impact on our model. CLDD (cooling degree days), AWND (average wind speed), HTDD (heating degree days), and population are features that have significant impact on the regression model. The interpretation of the coefficients is that, take the HTDD for example, the 862.7721 coefficient means that with an increase of 1 heating degree days, there will be an increase of 862.77215 million kilowatt hours of sales of electricity.

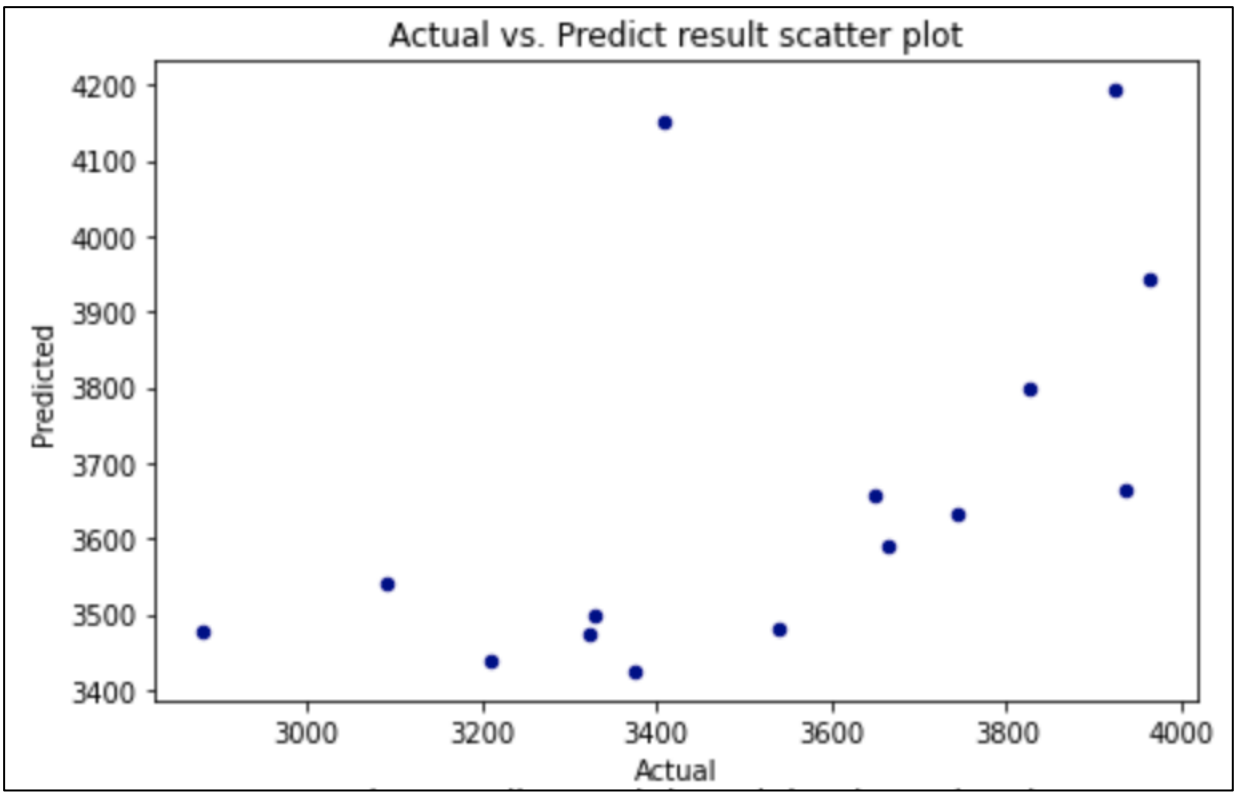
MAE, MSE, RMSE

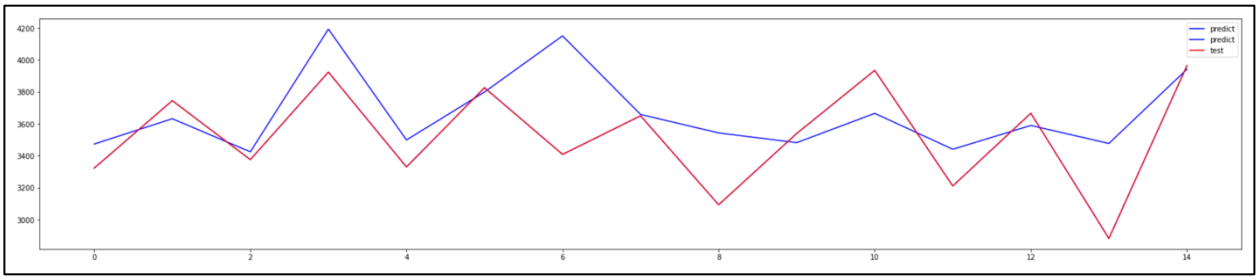
```
Mean Absolute Error: 215.5343495979653
Mean Squared Error: 92342.39217748676
Root Mean Squared Error: 303.87891038617136
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Prediction visualization







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted. In all three graphs, we can tell that the model is not representing the relationship between the features that we select and the retail sales of electricity that we want to predict, so in the ANN part, we will explore different methods to find a better representation for commercial sector data.

1.2 Industrial Sector for Pennsylvania

1.2.1 VIF

```
Cents/kWh      1.980803e+00
CLDD           1.378847e+01
AWND           3.725321e+00
TAVG           3.940180e+02
area           4.481568e+06
HTDD           2.983458e+02
Population     1.547828e+00
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove TAVG and area to reduce the multicollinearity from the dataset.

```
const          4.347858e+06
Cents/kWh      1.878793e+00
CLDD           3.014123e+00
AWND           3.721805e+00
HTDD           4.252958e+00
Population     1.519247e+00
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

1.2.2 Regression on the Industrial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	industrial_usage	R-squared:	0.666
Model:	OLS	Adj. R-squared:	0.602
Method:	Least Squares	F-statistic:	10.40
Date:	Wed, 22 Jul 2020	Prob (F-statistic):	1.27e-08
Time:	13:41:10	Log-Likelihood:	-355.02
No. Observations:	57	AIC:	730.0
Df Residuals:	47	BIC:	750.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3754.4942	141.476	26.538	0.000	3469.881	4039.107
Cents/kWh	61.8447	142.145	0.435	0.665	-224.115	347.805
CLDD	351.2713	158.556	2.215	0.032	32.298	670.244
AWND	-332.5253	169.693	-1.960	0.056	-673.903	8.853
HTDD	240.0498	210.078	1.143	0.259	-182.572	662.672
Population	222.2296	61.797	3.596	0.001	97.910	346.549
solar-generation	176.0024	150.702	1.168	0.249	-127.170	479.175
Season_Fall	130.6492	81.145	1.610	0.114	-32.593	293.892
Season_Summer	7.4475	87.949	0.085	0.933	-169.483	184.378
Season_Winter	69.3232	92.047	0.753	0.455	-115.851	254.497

Omnibus:	0.329	Durbin-Watson:	1.860
Prob(Omnibus):	0.848	Jarque-Bera (JB):	0.397
Skew:	0.169	Prob(JB):	0.820
Kurtosis:	2.770	Cond. No.	22.2

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.666, and the adjusted R-squared is 0.602, there is not much difference in them, so we are safe to say that our

input variables are all valuable to the model, and they can explain 66.6% of the overall result.

Feature coefficients and Significance

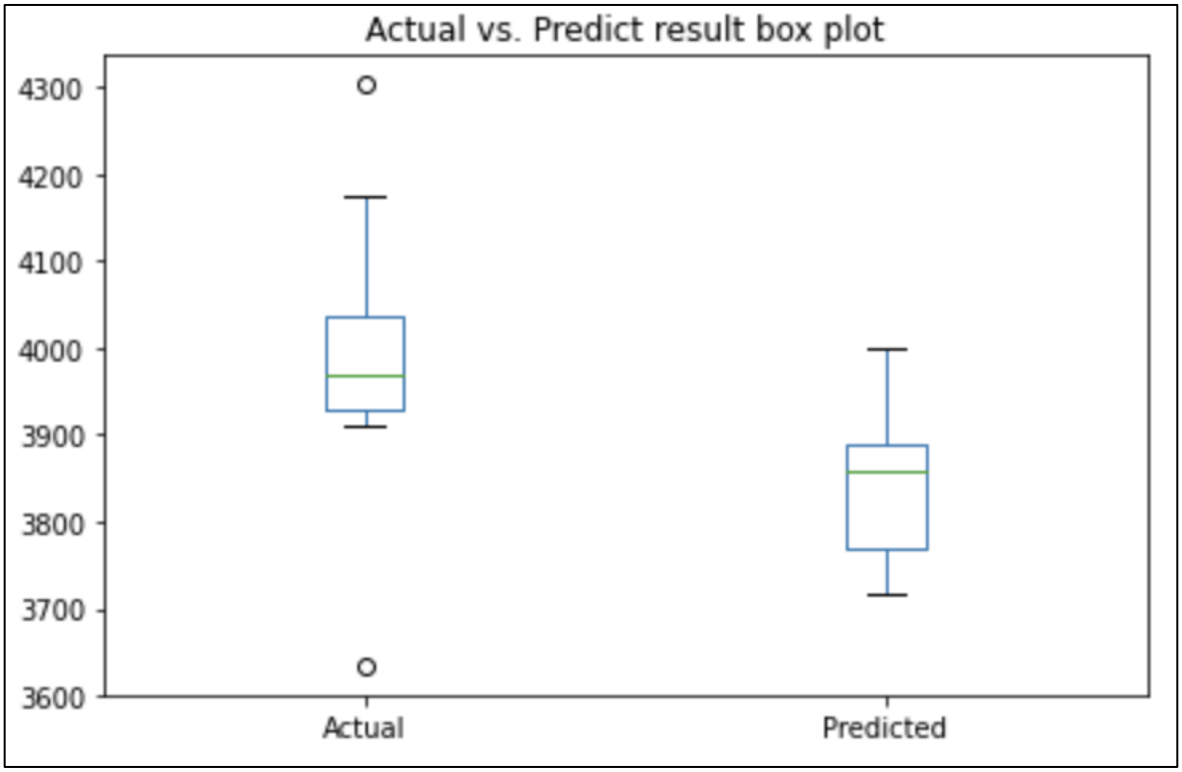
From the above table, we can say that CLDD and population are important features in our model. The interpretation is that with one more day of cooling degree days, we are expecting an increase of 351.2713 mkWh in electricity consumption.

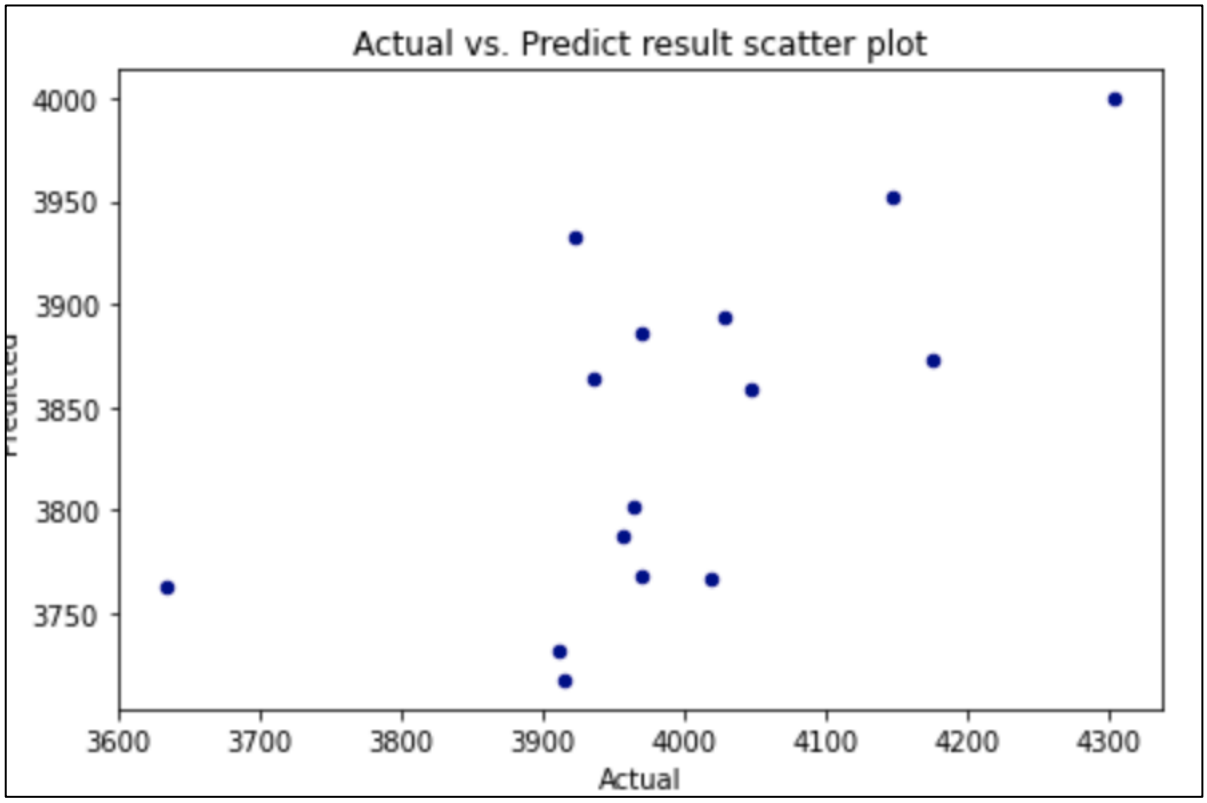
MAE, MSE, RMSE

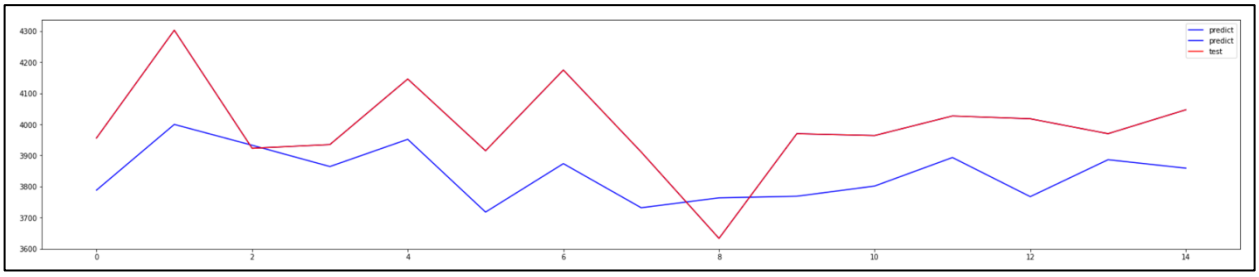
```
Mean Absolute Error: 171.630089969333
Mean Squared Error: 35464.04925772611
Root Mean Squared Error: 188.31900928405
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Prediction visualization







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted. In all three graphs, we can tell that the model is not representing the relationship between the features that we select and the retail sales of electricity that we want to predict, so in the ANN part, we will explore different methods to find a better representation for commercial sector data.

1.3 Residential Sector for Pennsylvania

1.3.1 VIF

```
Cents/kWh      1.836973e+00
CLDD           1.295634e+01
AWND           3.673565e+00
HTDD           2.952320e+02
TAVG           3.755230e+02
area           3.056286e+06
Population     1.095045e+00
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove TAVG and area to reduce the multicollinearity from the dataset.

```
const          3.054214e+06
Cents/kWh      1.828184e+00
CLDD           3.388102e+00
AWND           3.661130e+00
HTDD           4.706669e+00
Population     1.094681e+00
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

1.3.2 Regression on the residential sector

We used the OLS model to do the regression, and the output of the regression is as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      residential_usage   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.852
Method:                 Least Squares   F-statistic:                     36.72
Date:                Mon, 20 Jul 2020   Prob (F-statistic):           2.60e-18
Time:                        10:45:05   Log-Likelihood:                -399.25
No. Observations:                  57   AIC:                             818.5
Df Residuals:                      47   BIC:                             838.9
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.287e+04   1.52e+05      0.085      0.933   -2.92e+05    3.18e+05
Season_Summer    -80.7448    187.303     -0.431      0.668    -457.550     296.060
Season_Fall     -143.6936    154.867     -0.928      0.358    -455.246     167.859
Season_Winter    297.6897    172.688      1.724      0.091     -49.715     645.094
Cents/kWh      -4702.8595   1885.853     -2.494      0.016   -8496.707    -909.012
CLDD            2441.4608    352.685      6.922      0.000    1731.950    3150.972
AWND           -1454.2813    707.897     -2.054      0.046   -2878.387     -30.175
HTDD            2759.5682    409.618      6.737      0.000    1935.524    3583.613
Population     -4517.6116   1.52e+05     -0.030      0.976    -3.1e+05    3.01e+05
solar-generation   290.2957    481.796      0.603      0.550    -678.953    1259.545
==============================================================================
Omnibus:                        1.939   Durbin-Watson:                   1.549
Prob(Omnibus):                  0.379   Jarque-Bera (JB):                1.315
Skew:                           0.359   Prob(JB):                        0.518
Kurtosis:                       3.192   Cond. No.                     1.10e+04
==============================================================================
```

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.875, and the adjusted R-squared is 0.852, there is not much difference in them, so we are safe to say that our

input variables are all valuable to the model, and they can explain 87.5% of the overall result.

Feature coefficients and Significance

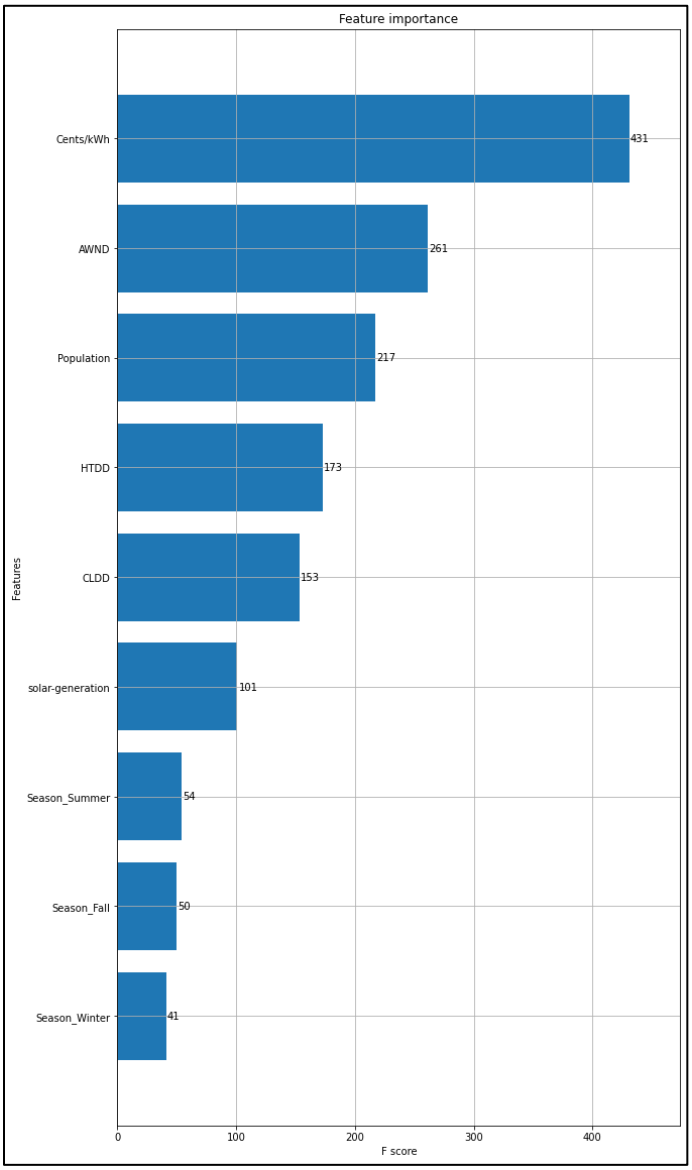
From the above table, we can identify some features that have significant impact on our model. The price of electricity (Cents/kWh), CLDD (cooling degree days), AWND (average wind speed), HTDD (heating degree days), are features that have significant impact on the regression model. The interpretation of the coefficients is that, take the Cents/kWh for example, the -4702.8595 coefficient means that with an increase of 1 cent in electricity price, there will be a decrease of 4702.8595 million kilowatt hours of sales of electricity.

MAE, MSE, RMSE

```
Mean Absolute Error: 202.7552706066282
Mean Squared Error: 63779.46390997826
Root Mean Squared Error: 252.54596395503583
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

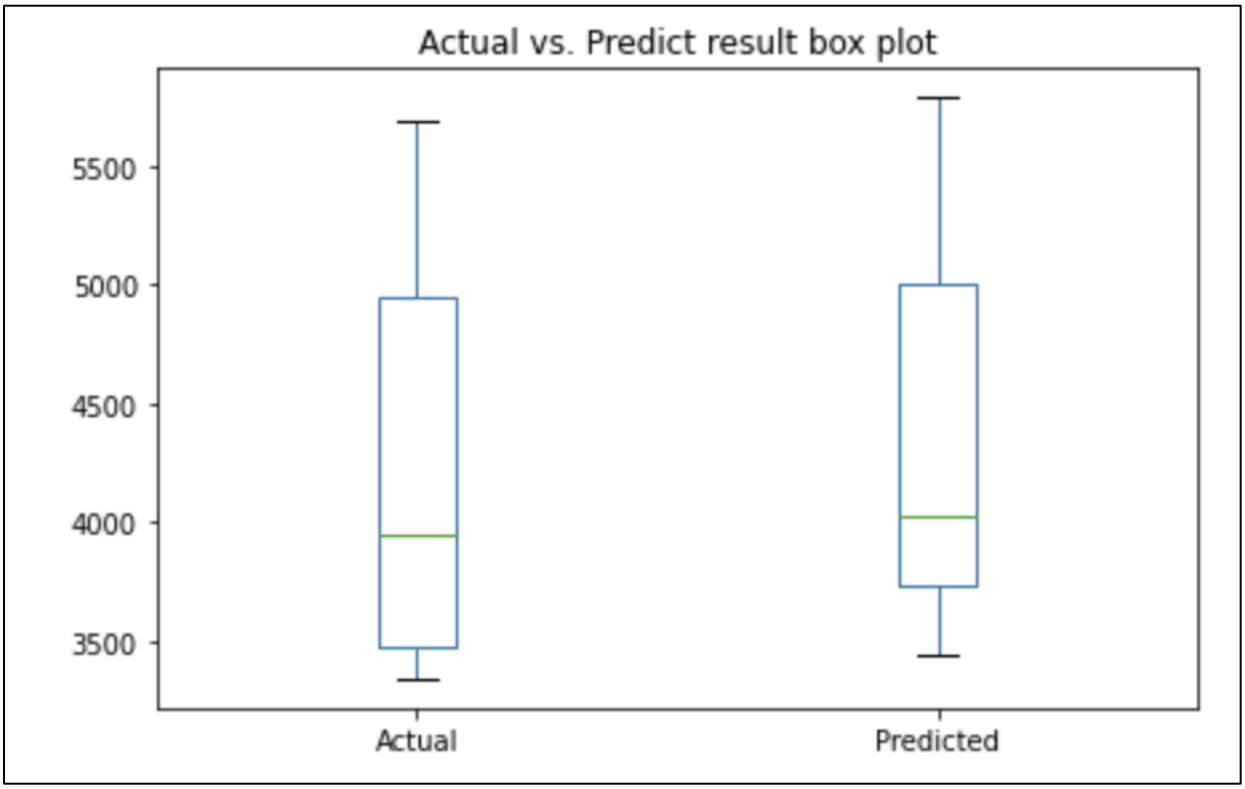
Feature Importance

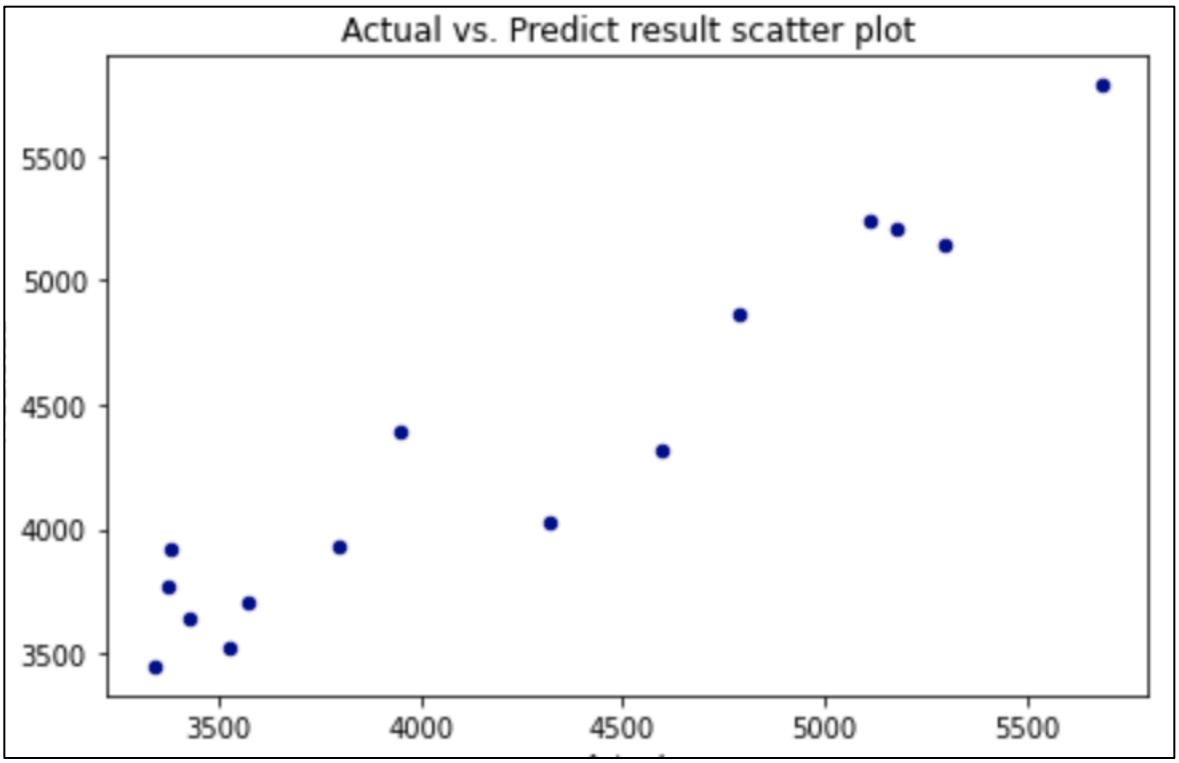


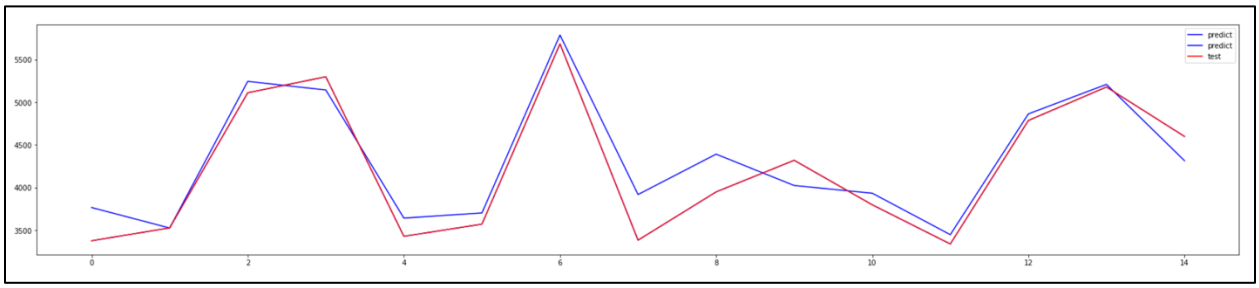
Using the feature importance function in linear models, we can identify features that are playing an important role in the model. In the residential sector, the price of electricity is the most important feature. This is quite intuitive, since residents are most sensitive to the price of electricity when they are making consumption.

The average wind day and population ranks second and third. This wind day may seem not directly related, but the population is surely the obvious factor influencing electricity consumption.

Actual vs Prediction visualization







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted. In all three graphs, we can tell that the model is representing the relationship between the features that we select and the retail sales of electricity that we want to predict.

2. North Carolina

2.1 Commercial Sector for North Carolina

Retail Price(Cents/kWh)	1.275370
CLDD	111.581829
AWND	1.600985
TAVG	839.574851
area	24470.154837
HTDD	402.342286
population	1.190347
dtype: float64	

2.1.1 VIF

When we put all the variables into the model, we can see that area, TAVG, HTDD, CLDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove TAVG and area to reduce the multicollinearity from the dataset.

```
const                          4395.888096
Retail Price(Cents/kWh)           1.178066
CLDD                              3.208894
AWND                              1.557398
HTDD                              2.735452
population                        1.166792
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

2.1.2 Regression on the commercial sector

We used the OLS model to do the regression, and the output of the regression is

OLS Regression Results

Dep. Variable:	mkwh	R-squared:	0.942
Model:	OLS	Adj. R-squared:	0.931
Method:	Least Squares	F-statistic:	85.47
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	4.48e-26
Time:	12:31:11	Log-Likelihood:	-346.40
No. Observations:	57	AIC:	712.8
Df Residuals:	47	BIC:	733.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3414.5879	1452.384	2.351	0.023	492.768	6336.407
Summer	-21.3839	87.147	-0.245	0.807	-196.702	153.934
Fall	134.4376	65.138	2.064	0.045	3.397	265.478
Winter	93.1895	79.698	1.169	0.248	-67.142	253.521
Retail Price(Cents/kWh)	-212.3245	665.436	-0.319	0.751	-1551.009	1126.360
CLDD	1569.4179	140.275	11.188	0.000	1287.221	1851.615
AWND	-65.6538	232.544	-0.282	0.779	-533.472	402.165
HTDD	388.7909	142.107	2.736	0.009	102.909	674.673
population	193.0307	1551.048	0.124	0.901	-2927.276	3313.338
solar-generation	82.9808	174.040	0.477	0.636	-267.143	433.105

Omnibus:	0.430	Durbin-Watson:	1.897
Prob(Omnibus):	0.807	Jarque-Bera (JB):	0.407
Skew:	0.191	Prob(JB):	0.816
Kurtosis:	2.841	Cond. No.	280.

as follows:

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.942, and the adjusted R-squared is 0.931, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 94.2% of the overall result.

Feature coefficients and Significance

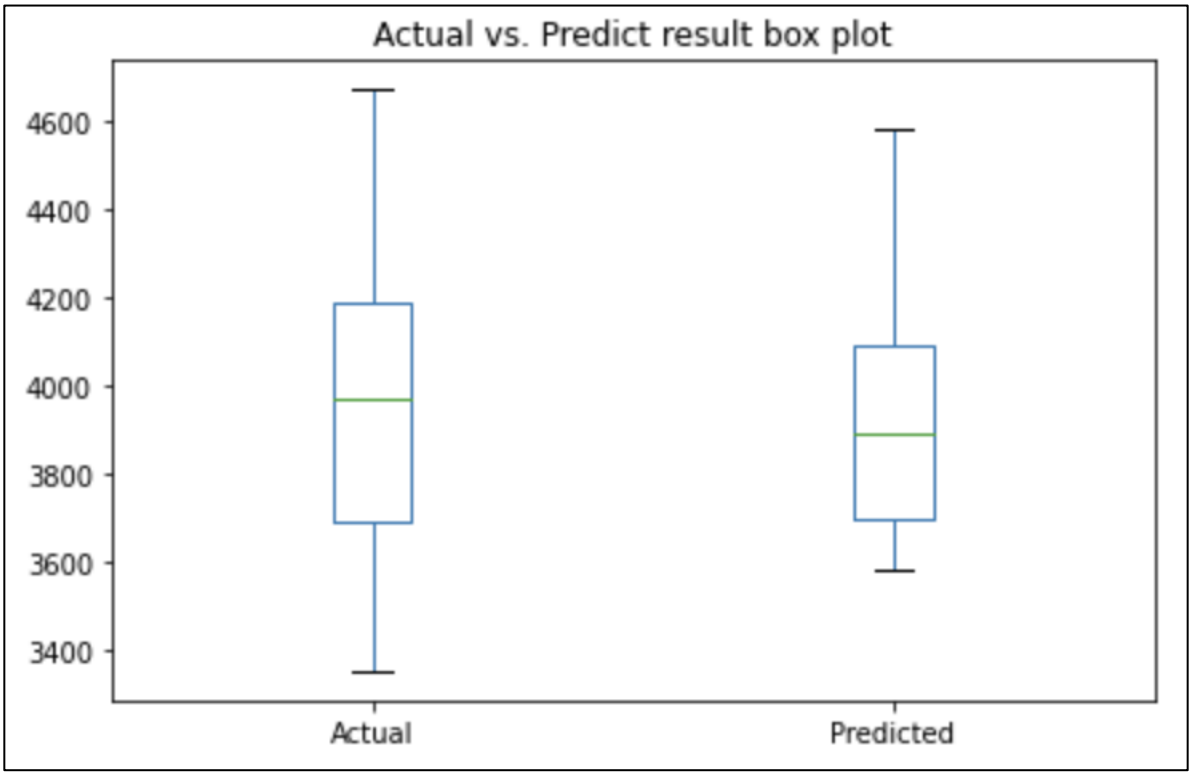
From the above table, we can only say that CLDD and HTDD is a good indicator of electricity consumption in the commercial sector. The logic is that when there is 1 day increase of cooling degree days, there will be a corresponding 1569.4179 mkWh increase in electricity consumption.

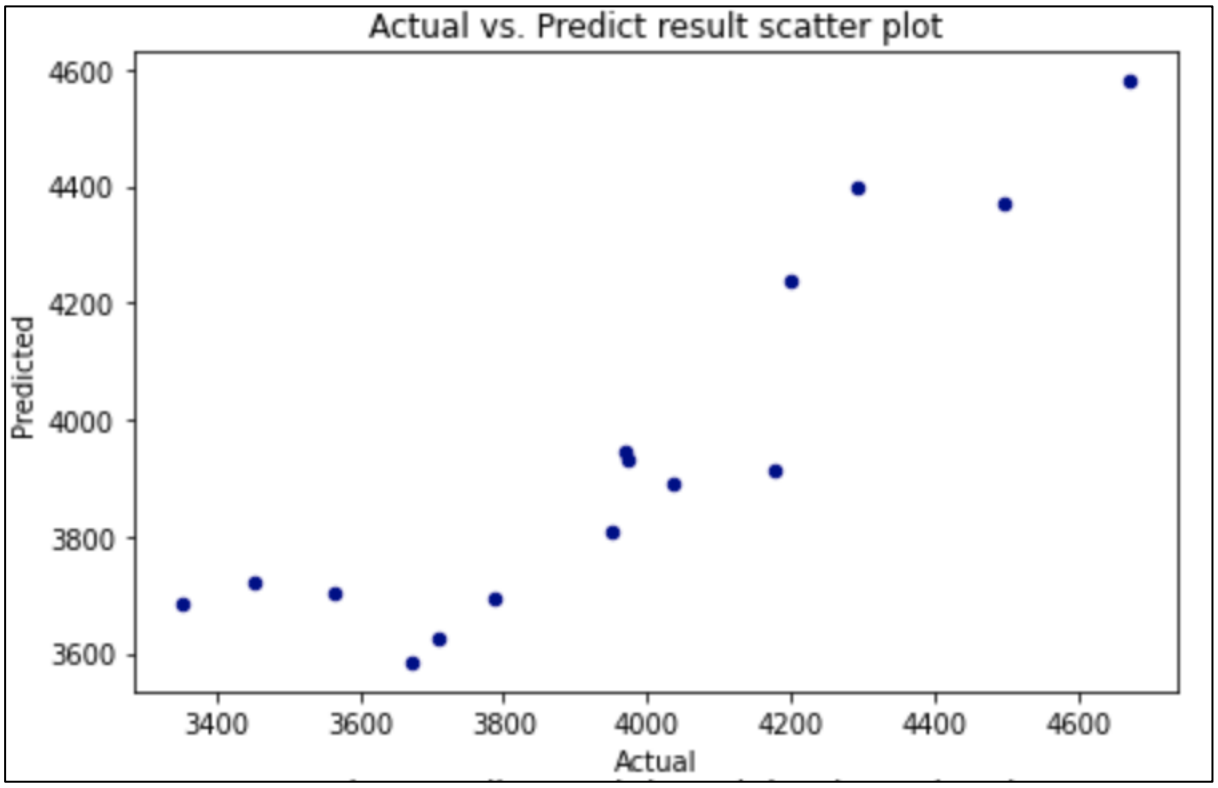
MAE, MSE, RMSE

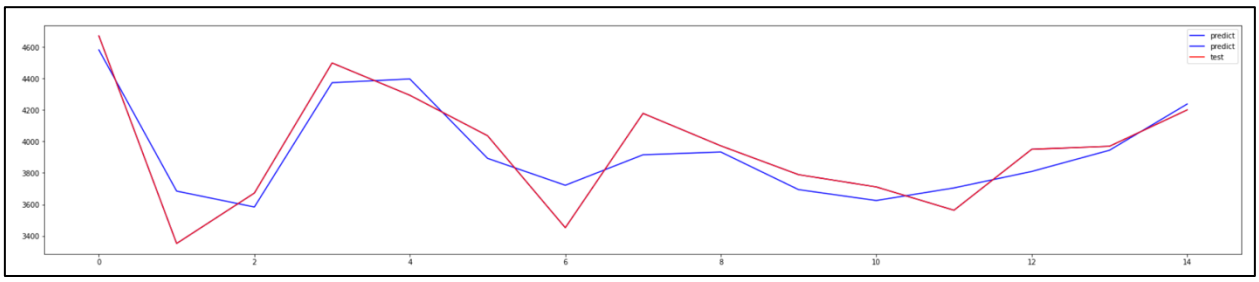
```
Mean Absolute Error: 132.01828160005192
Mean Squared Error: 25005.032319111375
Root Mean Squared Error: 158.12979579798164
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Prediction:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted. In all three graphs, we can tell that the model is representing the relationship between the features that we select and the retail sales of electricity that we want to predict. The

trend in both graphs shows that the model that we implemented is a good representation of the scenario. The high error in MAE and MSE may be due to some extreme values that are not representable by linear models.

2.2 Regression on Industrial Sector for North Carolina

2.2.1 VIF

```
Cents/kWh          2.583849
CLDD             105.036736
AWND               1.601164
HTDD             370.549805
Population         1.438008
TAVG             777.178386
area           21558.240577
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD, CLDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove TAVG and area to reduce the multicollinearity from the dataset.

```
const          5350.720888
Cents/kWh         2.578334
CLDD              5.289405
AWND              1.547069
HTDD              2.910134
Population        1.415069
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

2.2.2 Regression on the Industrial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	industrial_usage	R-squared:	0.635
Model:	OLS	Adj. R-squared:	0.565
Method:	Least Squares	F-statistic:	9.066
Date:	Sun, 19 Jul 2020	Prob (F-statistic):	8.77e-08
Time:	17:16:08	Log-Likelihood:	-343.61
No. Observations:	57	AIC:	707.2
Df Residuals:	47	BIC:	727.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4900.7665	1497.146	3.273	0.002	1888.896	7912.637
Season_Summer	-64.6737	88.555	-0.730	0.469	-242.823	113.476
Season_Fall	-112.9645	67.192	-1.681	0.099	-248.137	22.208
Season_Winter	-244.6516	81.028	-3.019	0.004	-407.659	-81.644
Cents/kWh	-392.4867	531.286	-0.739	0.464	-1461.296	676.322
CLDD	236.1124	141.123	1.673	0.101	-47.791	520.016
AWND	-660.9808	264.699	-2.497	0.016	-1193.487	-128.474
HTDD	106.6315	127.309	0.838	0.407	-149.481	362.744
Population	-1887.7154	1365.737	-1.382	0.173	-4635.223	859.792
solar-generation	87.4061	50.662	1.725	0.091	-14.512	189.324

Omnibus:	1.481	Durbin-Watson:	2.297
Prob(Omnibus):	0.477	Jarque-Bera (JB):	0.961
Skew:	0.310	Prob(JB):	0.618
Kurtosis:	3.142	Cond. No.	271.

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.635, and the adjusted R-squared is 0.565, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 63.5% of the overall result.

Feature coefficients and Significance

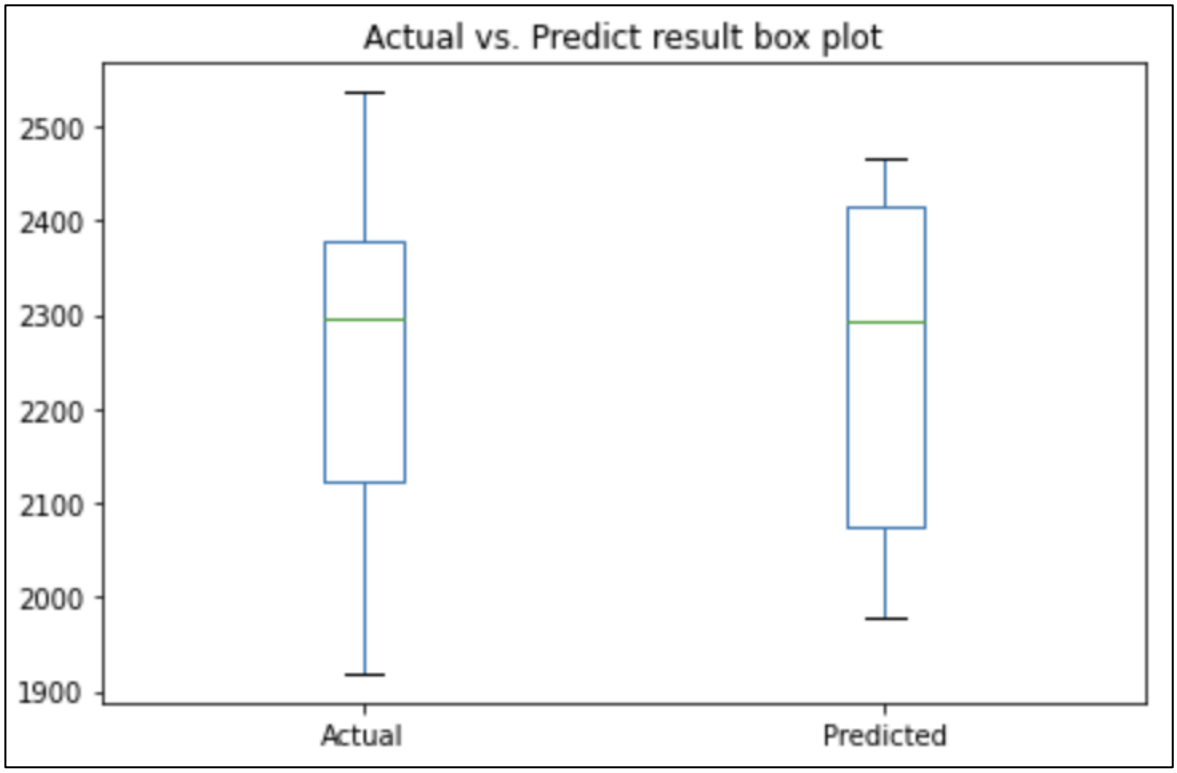
From the above table, we can only say that AWND and Winter is a good indicator of electricity consumption in the industrial sector. The logic is that since we choose spring as the baseline value, the industrial sector in North Carolina consumes 244.6516 mkWh less electricity in Winter than in Spring.

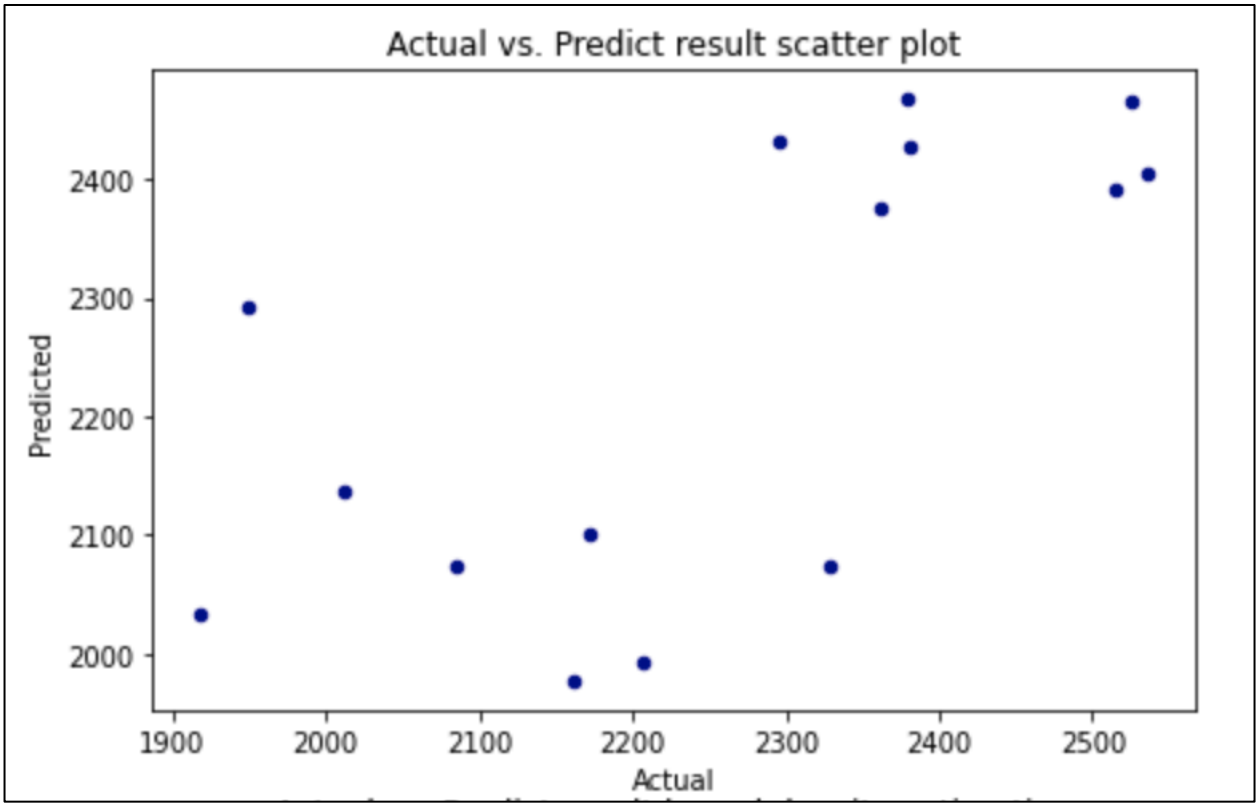
MAE, MSE, RMSE:

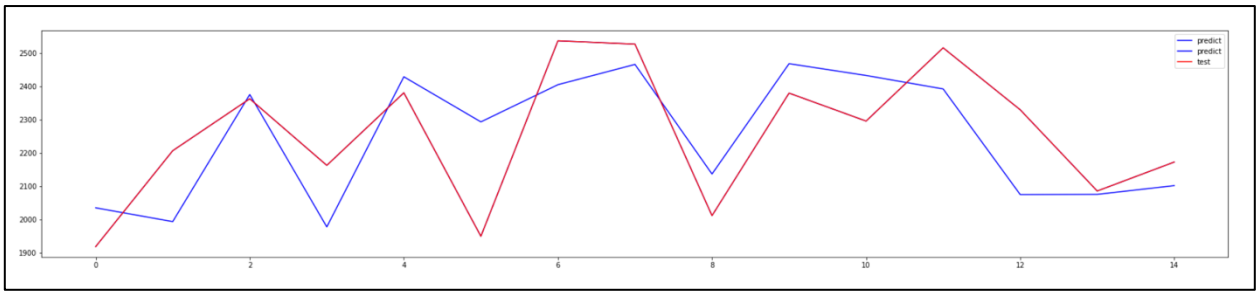
```
Mean Absolute Error: 128.04039434670065
Mean Squared Error: 24134.39316155226
Root Mean Squared Error: 155.3524803842934
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Prediction:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted. In all three graphs, we can tell that the model is not representing the relationship between the features that we select and the retail sales of electricity that we want to predict. The trend in both graphs shows that the model that we implemented is quite opposite from the real scenario. One explanation for this may be that there are other variables influencing the industrial sector rather than the ones we choose.

2.3 Residential Sector for North Carolina

2.3.1 VIF

```
Cents/kWh          3.050901
CLDD             102.068185
AWND               1.683081
HTDD             406.347948
Population         1.202919
TAVG             806.733599
area           22067.850300
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove TAVG and area to reduce the multicollinearity from the dataset.

```
const          3984.681908
Cents/kWh         2.932856
CLDD              4.219887
AWND              1.654523
HTDD              7.172047
Population        1.164622
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables

2.3.2 Regression on the residential sector

We used the OLS model to do the regression, and the output of the regression is

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      residential_usage   R-squared:                       0.920
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                     60.19
Date:                Sun, 19 Jul 2020   Prob (F-statistic):           8.94e-23
Time:                        17:04:58   Log-Likelihood:                -399.96
No. Observations:                  57   AIC:                             819.9
Df Residuals:                      47   BIC:                             840.4
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        351.6744   9062.499      0.039      0.969   -1.79e+04    1.86e+04
Season_Summer -161.9200    218.718     -0.740      0.463    -601.924     278.084
Season_Fall  -181.8732    171.023     -1.063      0.293    -525.927     162.180
Season_Winter  324.5047    188.332      1.723      0.091     -54.370     703.379
Cents/kWh    1284.4463   2157.218      0.595      0.554   -3055.316    5624.209
CLDD         4489.8310    368.446     12.186      0.000    3748.614    5231.048
AWND         -285.4262    598.788     -0.477      0.636   -1490.032     919.179
HTDD         4000.9692    439.362      9.106      0.000    3117.087    4884.851
Population   1152.0544   8423.038      0.137      0.892   -1.58e+04    1.81e+04
solar-generation  -94.5262    607.445     -0.156      0.877   -1316.547    1127.495
==============================================================================
Omnibus:                        7.655   Durbin-Watson:                   2.269
Prob(Omnibus):                  0.022   Jarque-Bera (JB):                7.593
Skew:                           0.606   Prob(JB):                       0.0224
Kurtosis:                       4.315   Cond. No.                         617.
==============================================================================
```

as follows:

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.92, and the adjusted R-squared is 0.905, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 92% of the overall result.

Feature coefficients and Significance

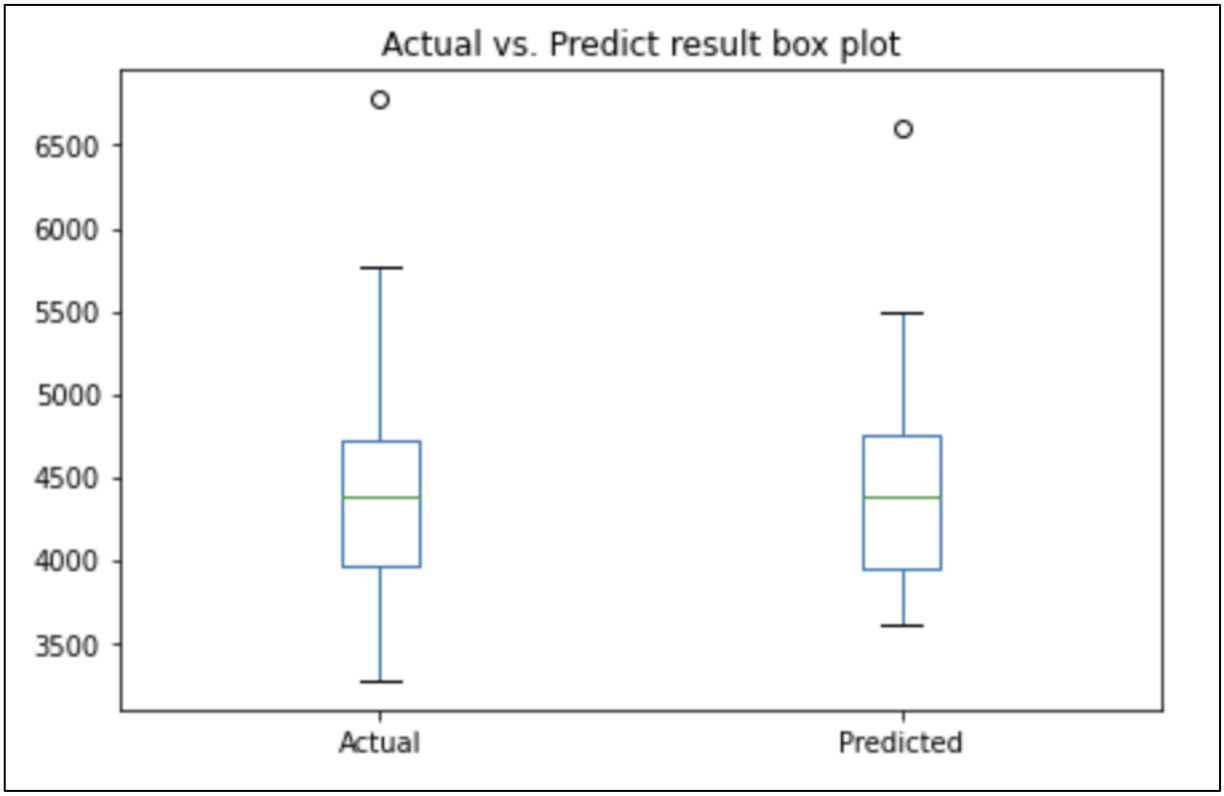
From the above table, we can only say that CLDD and HTDD is a good indicator of electricity consumption in the residential sector. The logic is that when there is 1 day increase of cooling degree days, there will be a corresponding 4489.8310 mkWh increase in electricity consumption.

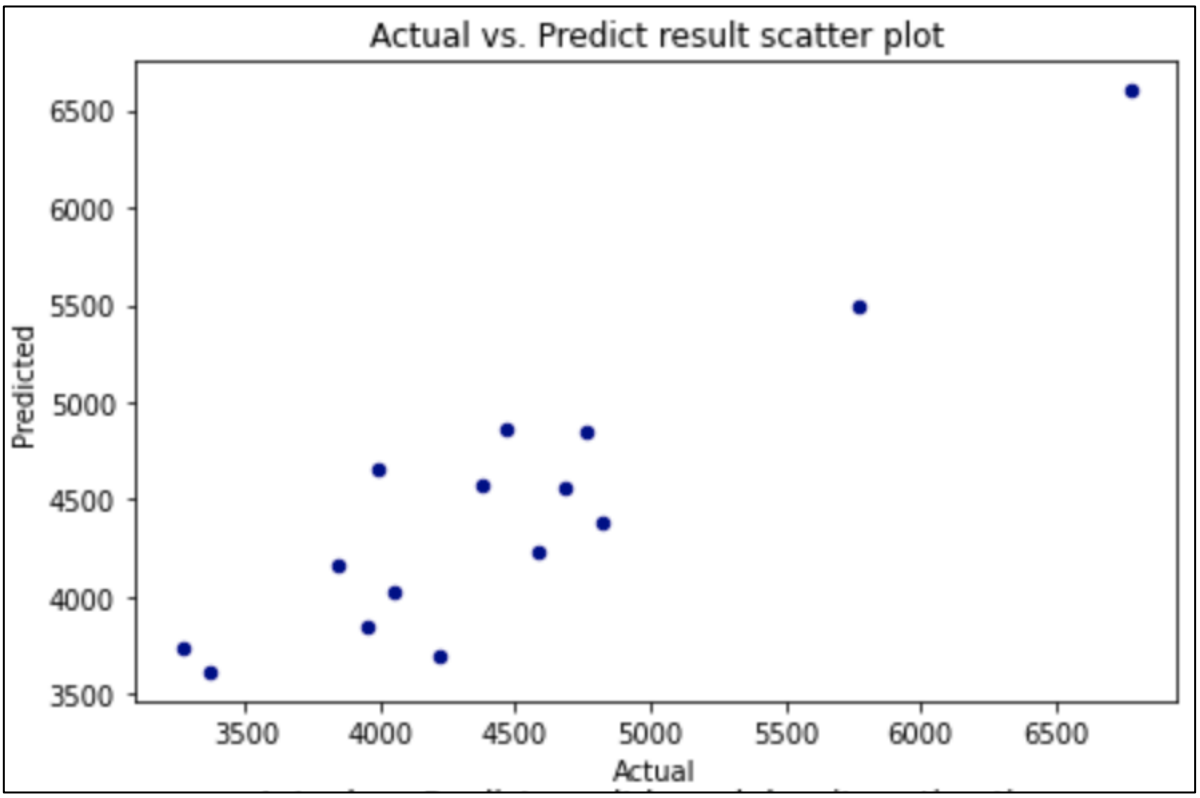
MAE, MSE, RMSE

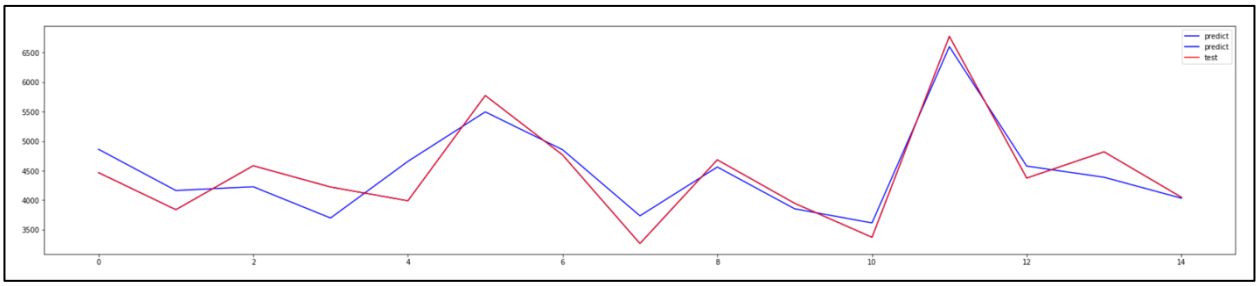
```
Mean Absolute Error: 292.6548506506423
Mean Squared Error: 117185.508712844
Root Mean Squared Error: 342.32368996732316
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Prediction:




Actual vs. Predict result scatter plot
Predicted
Actual
6500
6000
5500
5000
4500
4000
3500
3500
4000
4500
5000
5500
6000
6500


predict
predict
test

In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted. In all three graphs, we can tell that the model is representing the relationship between the features that we select and the retail sales of electricity that we want to predict. The high error in MAE and MSE may be due to some extreme values that are not representable by linear models.

3. Minnesota

3.1 Commercial Sector for Minnesota

3.1.1 VIF

Commercial_Retail Price	3.434971
CLDD	4.872860
TAVG	324.756687
AWND	2.719494
HTDD	302.680244
area	17359.252347
population	2.925733
solar-generation	3.103349
dtype: float64	

When we put all the variables into the model, we can see that TAVG, HTDD and area have large VIF, which means that the variables are multicollinear, as a result, we try to remove area to reduce the multicollinearity from the dataset.

```
const                          16074.729171
Commercial_Retail Price            3.242080
CLDD                               3.167480
AWND                               2.710098
HTDD                               3.539765
population                         2.807539
solar-generation                   3.086564
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

3.1.2 Regression on the Commercial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	Commercial_Usage	R-squared:	0.762
Model:	OLS	Adj. R-squared:	0.716
Method:	Least Squares	F-statistic:	16.68
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	6.97e-12
Time:	11:43:41	Log-Likelihood:	-317.08
No. Observations:	57	AIC:	654.2
Df Residuals:	47	BIC:	674.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1933.6490	67.109	28.813	0.000	1798.643	2068.655
Commercial_Retail Price	-107.5946	72.985	-1.474	0.147	-254.421	39.232
CLDD	343.5570	70.707	4.859	0.000	201.313	485.801
AWND	-182.0970	71.162	-2.559	0.014	-325.256	-38.938
HTDD	139.4681	68.594	2.033	0.048	1.476	277.461
population	-65.7432	52.544	-1.251	0.217	-171.447	39.961
solar-generation	65.1989	81.635	0.799	0.429	-99.029	229.427
Summer	39.0615	45.123	0.866	0.391	-51.715	129.838
Fall	11.6860	29.151	0.401	0.690	-46.958	70.330
Winter	10.6344	43.358	0.245	0.807	-76.591	97.860

Omnibus:	0.001	Durbin-Watson:	1.965
Prob(Omnibus):	0.999	Jarque-Bera (JB):	0.098
Skew:	0.009	Prob(JB):	0.952
Kurtosis:	2.798	Cond. No.	17.1

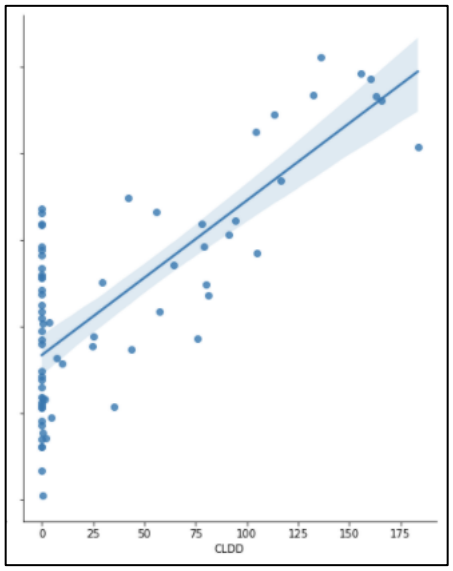
R-squared and Adj. R-squared:

From in our model, the R-squared is 0.762, and the adjusted R-squared is 0.716, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 76.2% of the overall result.

Feature coefficients and Significance

From the above table, we can say that CLDD, AWND and HTDD are good indicators of electricity consumption in the commercial sector. The logic is that for each increase in heating degree days, there will be a corresponding 139.4681 mkWh increase in electricity consumption.

Correlation Analysis:



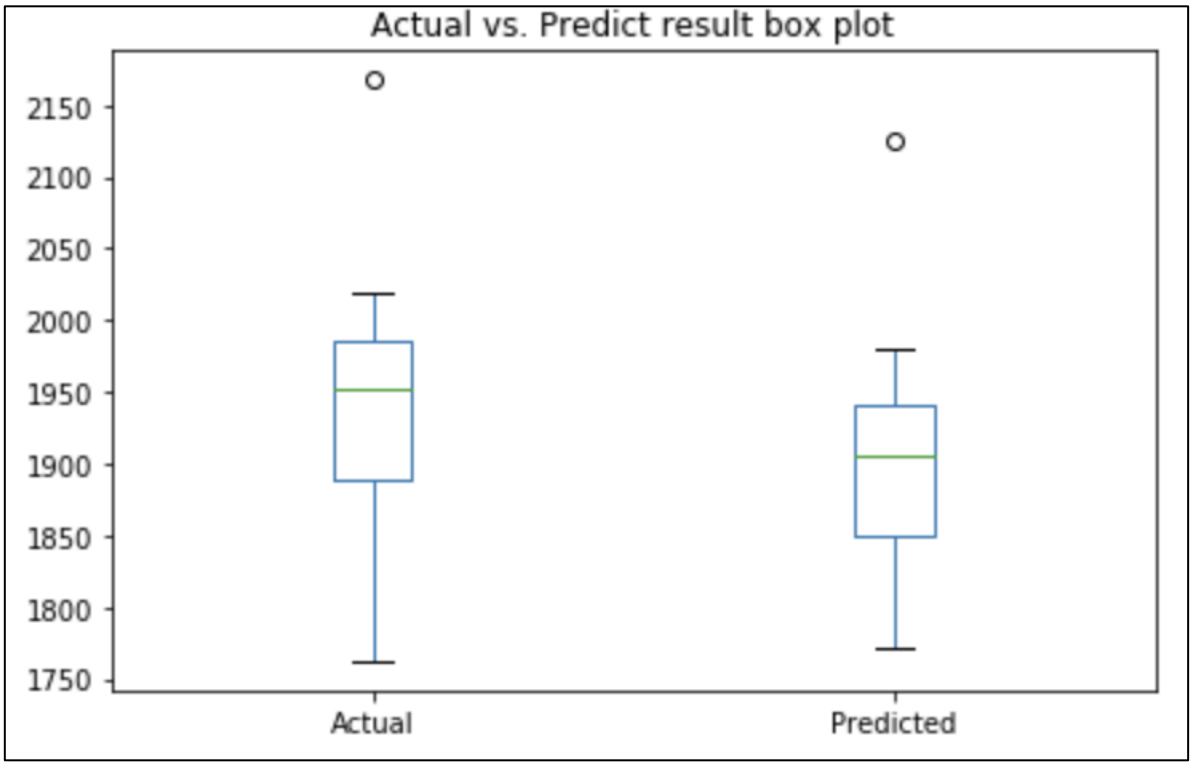
We can also tell from the correlation graph that cooling degree days is highly related to the electricity consumption.

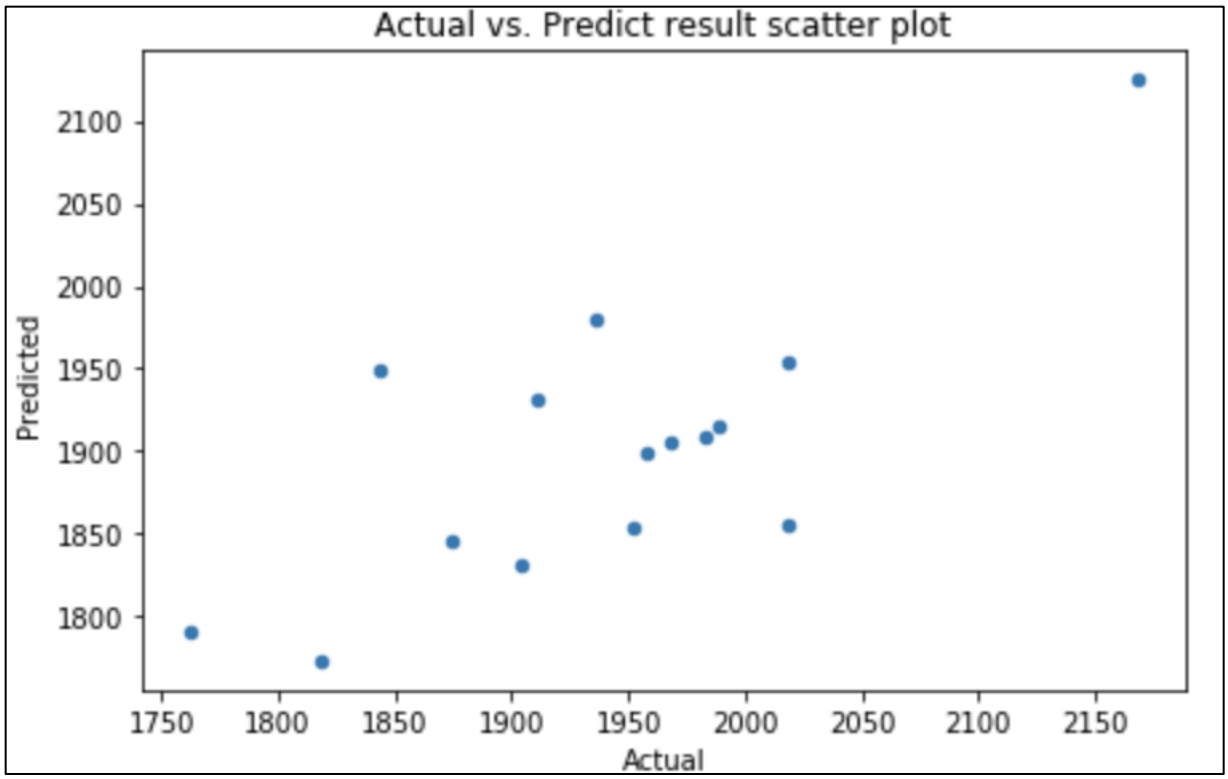
MAE, MSE, RMSE:

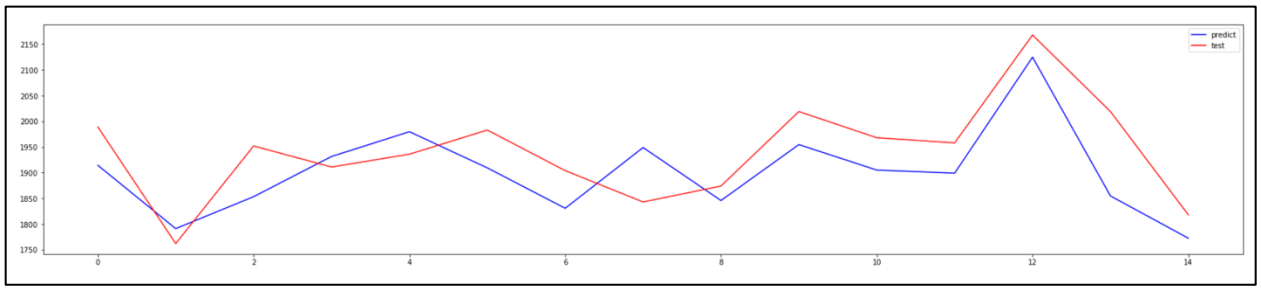
```
Performance Evaluation
Mean Absolute Error: 65.79307481977031
Mean Squared Error: 5586.5223558637435
Root Mean Squared Error: 74.74304219031858
```

The error metrics for the commercial sector in Minnesota has a relatively low error, so it is promising that the linear model here provides a good representation of the data.

Actual vs Predicted:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is in some way telling us the distribution of the data. However, our model cannot provide an accurate prediction of the actual electricity consumption.

3.2 Industrial Sector for Minnesota

3.2.1 VIF

Industrial_Retail Price	4.263805
CLDD	5.529707
TAVG	334.000492
AWND	2.700721
HTDD	308.267302
area	19847.143603
population	3.890219
solar-generation	3.120212
dtype: float64	

When we put all the variables into the model, we can see that area, HTDD, TAVG have large VIF, which means that the variables are multicollinear, as a result, we try to remove area to reduce the multicollinearity from the dataset.

const	19259.329871
Industrial_Retail Price	3.912992
CLDD	3.427203
AWND	2.698753
HTDD	3.086385
population	3.622430
solar-generation	3.109199
dtype: float64	

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

3.2.2 Regression on the Industrial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	Industrial_Usage	R-squared:	0.505
Model:	OLS	Adj. R-squared:	0.410
Method:	Least Squares	F-statistic:	5.320
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	5.30e-05
Time:	11:44:02	Log-Likelihood:	-329.08
No. Observations:	57	AIC:	678.2
Df Residuals:	47	BIC:	698.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1957.0245	76.430	25.605	0.000	1803.266	2110.783
Industrial_Retail Price	-279.7231	98.572	-2.838	0.007	-478.024	-81.423
CLDD	181.6630	88.713	2.048	0.046	3.196	360.130
AWND	-198.4315	86.082	-2.305	0.026	-371.606	-25.257
HTDD	101.5991	83.395	1.218	0.229	-66.171	269.369
population	-10.3627	78.362	-0.132	0.895	-168.006	147.280
solar-generation	68.7020	98.631	0.697	0.490	-129.719	267.123
Summer	16.7901	56.175	0.299	0.766	-96.220	129.800
Fall	31.5581	35.810	0.881	0.383	-40.483	103.599
Winter	-111.6324	53.741	-2.077	0.043	-219.745	-3.520

Omnibus:	2.832	Durbin-Watson:	1.791
Prob(Omnibus):	0.243	Jarque-Bera (JB):	2.167
Skew:	-0.323	Prob(JB):	0.338
Kurtosis:	2.297	Cond. No.	17.2

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.505, and the adjusted R-squared is 0.410, there is not much difference in them, so we are safe to say that our

input variables are all valuable to the model, and they can explain 50.5% of the overall result.

Feature coefficients and Significance

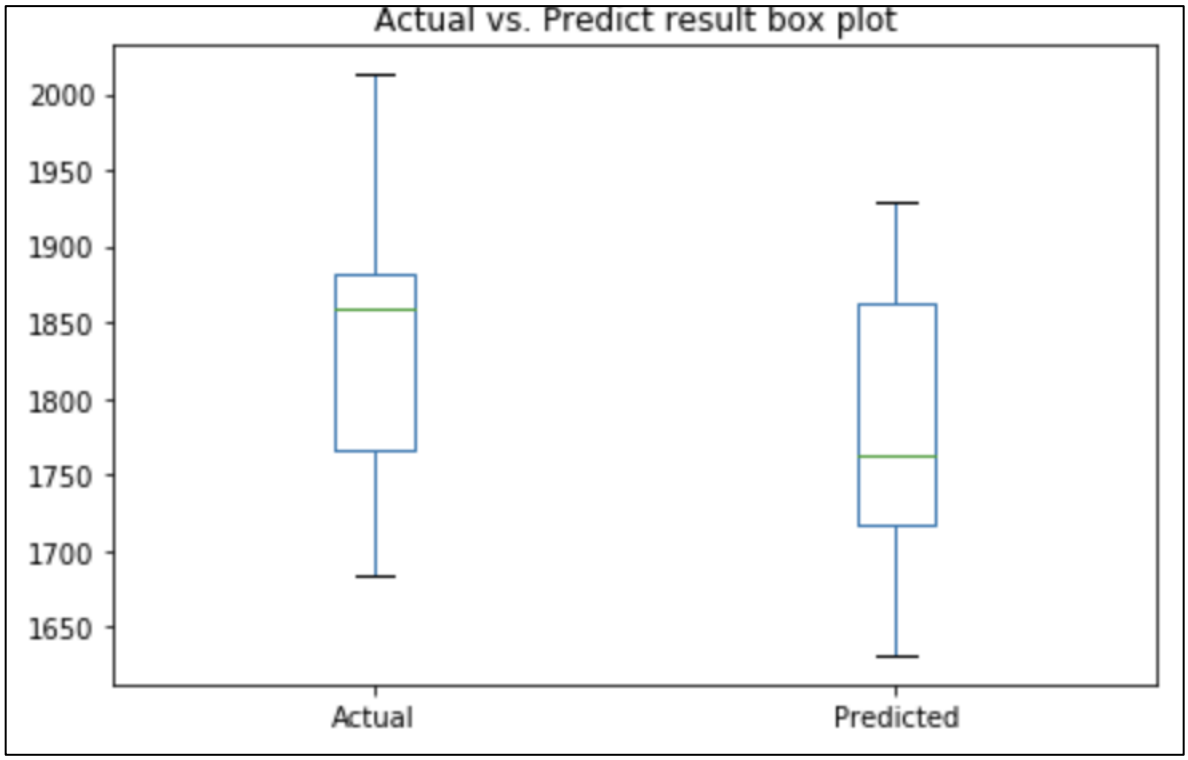
From the above table, we can say that price, CLDD, AWND and Winter season are significant features in the model. The logic is that with each increase in retail price of electricity, there will be a corresponding 279.7231 mkWh decrease in electricity consumption.

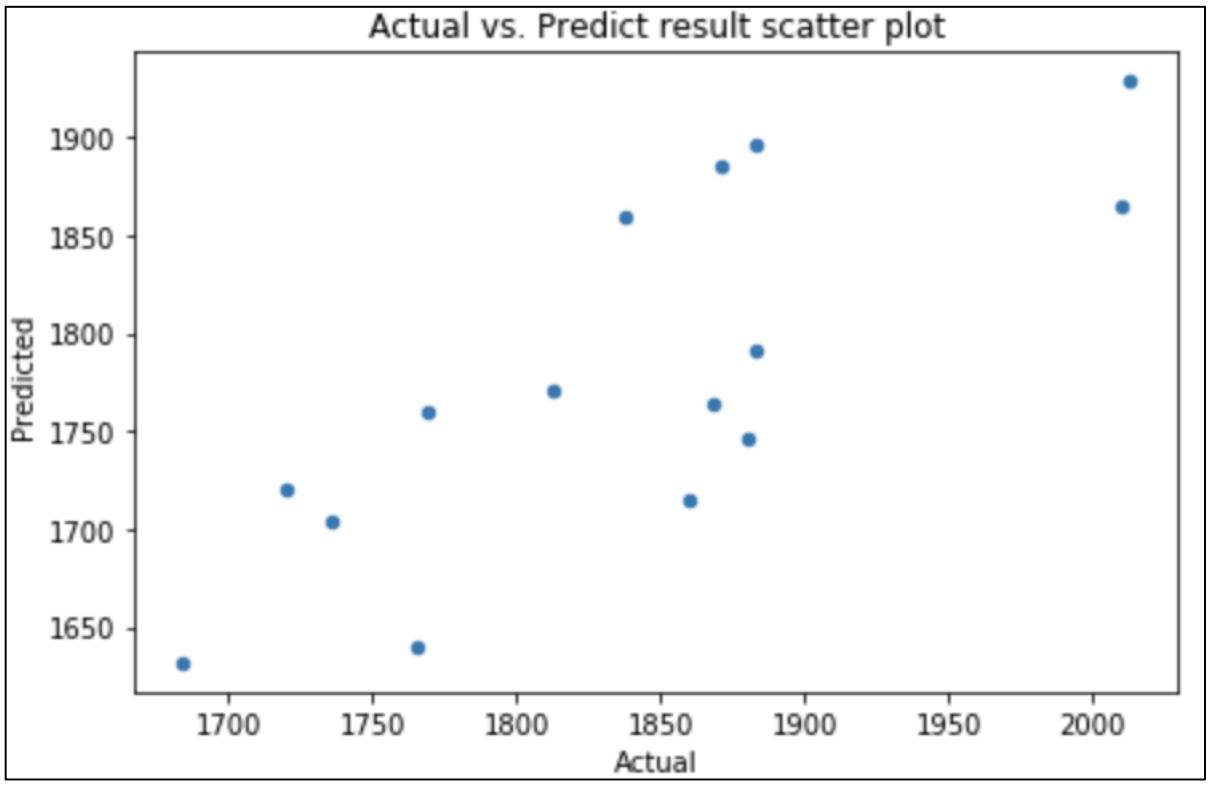
MAE, MSE, RMSE:

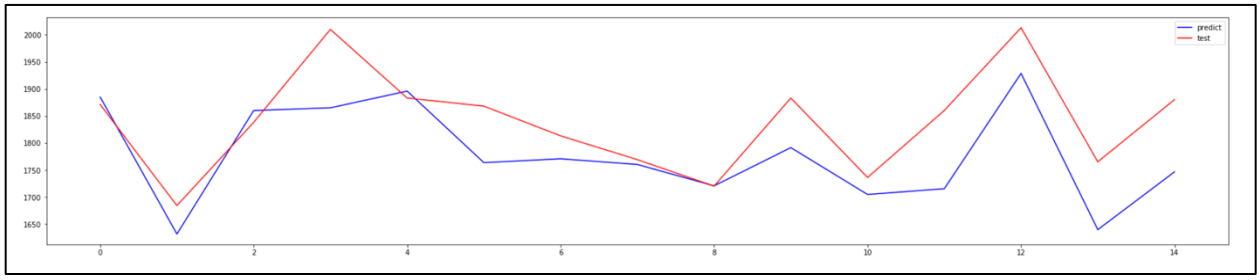
```
Performance Evaluation
Mean Absolute Error: 67.61194441108441
Mean Squared Error: 7238.424859798387
Root Mean Squared Error: 85.07893311389364
```

The error metrics for the industrial sector in Minnesota is low, which means that our linear model may be a good representation of the real data. We should compare with the figures to find out its real performance.

Actual vs Predicted:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is not a good representation of the relationship between the features that we select and the retail sales of electricity that we want to predict. The trend in both graphs are non-linear, and the values are quite far from each other. We may explore the ANN approach to better represent the data in the industrial sector.

3.3 Residential Sector for Minnesota

3.3.1 VIF

```
Residential_Retail Price        12.302592
CLDD                             4.692034
TAVG                           314.894587
AWND                             2.727114
HTDD                           305.419916
area                         26622.041385
population                       6.745918
solar-generation                 3.044430
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove area to reduce the multicollinearity from the dataset.

```
const                        26015.845152
Residential_Retail Price        11.975405
CLDD                             3.164633
AWND                             2.718872
HTDD                             8.500897
population                       6.488033
solar-generation                 3.012927
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

3.3.2 Regression on the Residential Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      residential_usage   R-squared:                       0.901
Model:                            OLS   Adj. R-squared:                  0.881
Method:                 Least Squares   F-statistic:                     47.28
Date:                Tue, 21 Jul 2020   Prob (F-statistic):           1.46e-20
Time:                        03:07:22   Log-Likelihood:                -331.32
No. Observations:                  57   AIC:                             682.6
Df Residuals:                      47   BIC:                             703.1
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const             1514.3496     85.595     17.692      0.000    1342.155    1686.545
Cents/kWh         -190.1938    125.902     -1.511      0.138    -443.476      63.089
CLDD               754.4757    104.268      7.236      0.000     544.715     964.236
AWND              -210.2111     85.061     -2.471      0.017    -381.332     -39.090
HTDD               834.3963     89.115      9.363      0.000     655.120    1013.672
Population          54.0333     95.281      0.567      0.573    -137.647     245.713
solar-generation   -32.0395     44.017     -0.728      0.470    -120.590      56.511
Season_Fall        -43.9027     34.652     -1.267      0.211    -113.613      25.808
Season_Summer       92.7229     65.987      1.405      0.167     -40.027     225.473
Season_Winter       93.8893     51.237      1.832      0.073      -9.186     196.964
==============================================================================
Omnibus:                        1.618   Durbin-Watson:                   2.279
Prob(Omnibus):                  0.445   Jarque-Bera (JB):                1.175
Skew:                           0.061   Prob(JB):                        0.556
Kurtosis:                       2.307   Cond. No.                         20.9
==============================================================================
```

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.901, and the adjusted R-squared is 0.881 there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 95% of the overall result.

Feature coefficients and Significance

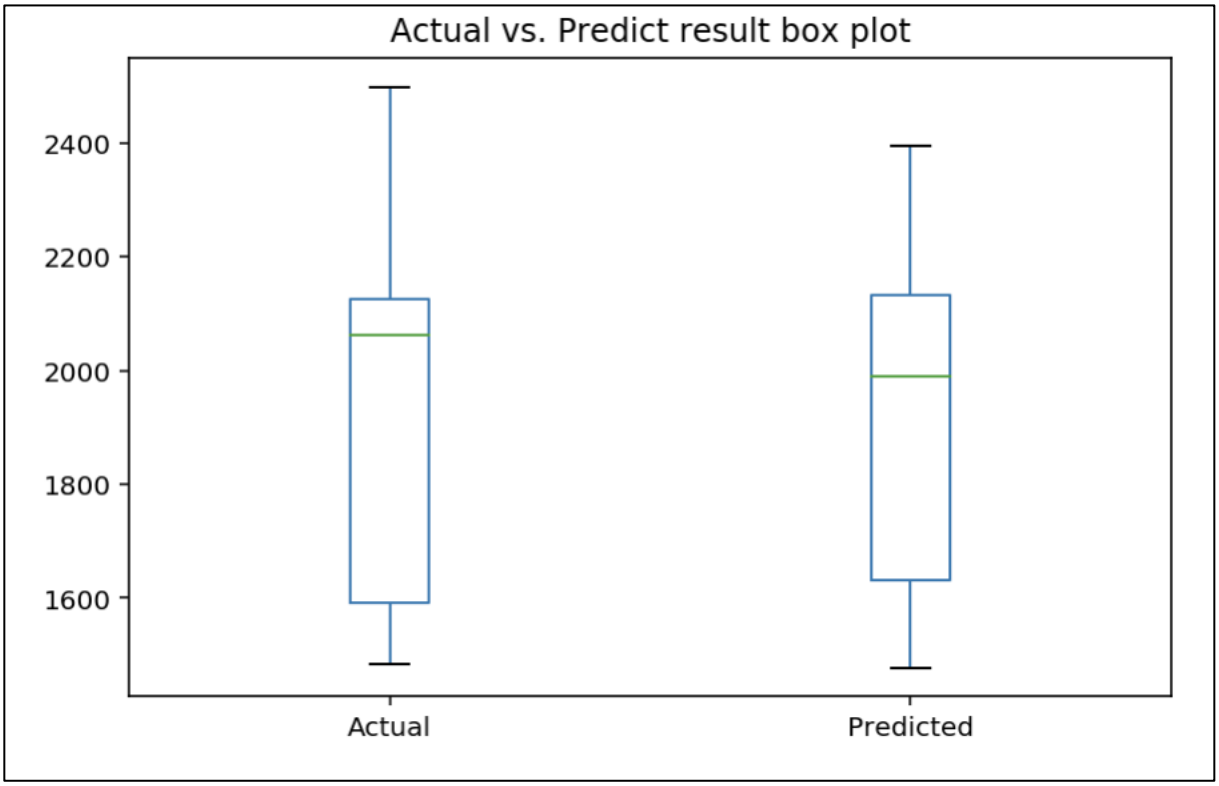
From the above table, we can say that CLDD and HTDD are good indicator of electricity consumption in the residential sector. The logic is that with each increase in cooling degree days, there will be a corresponding 754.4757 mkWh increase in electricity consumption.

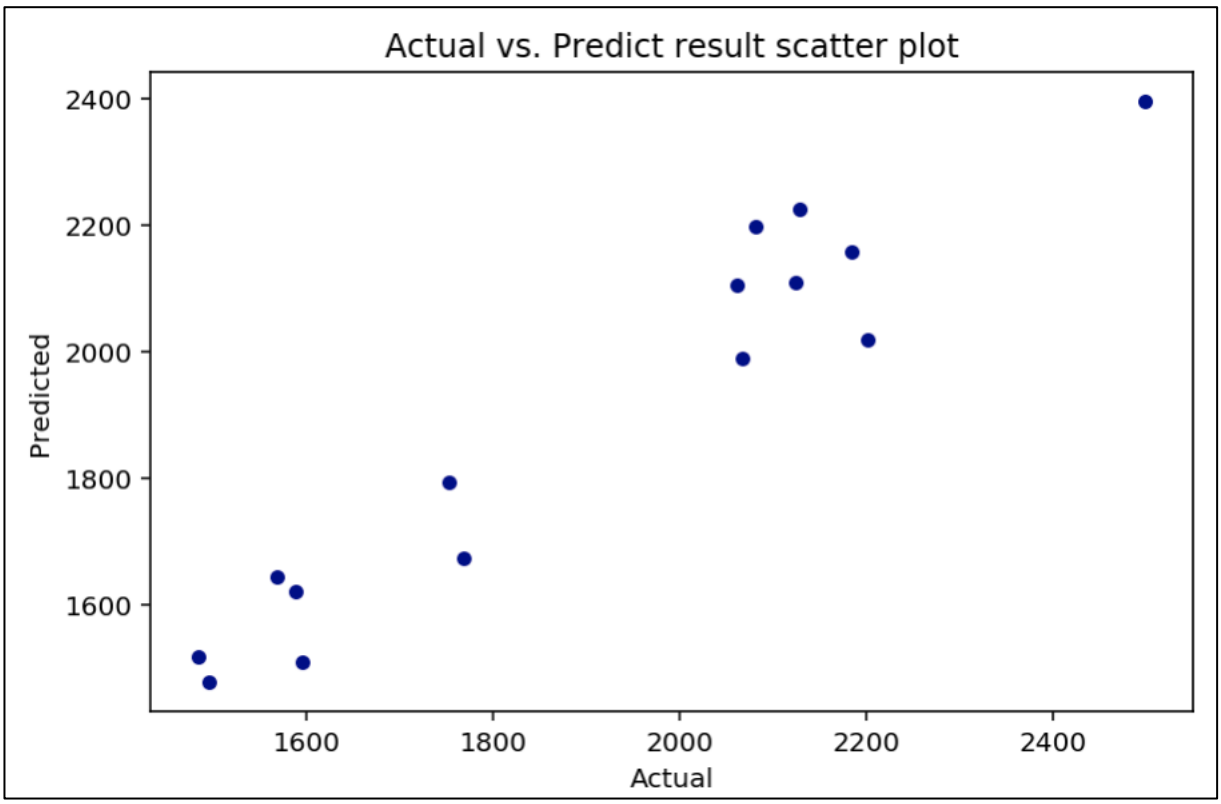
MAE, MSE, RMSE:

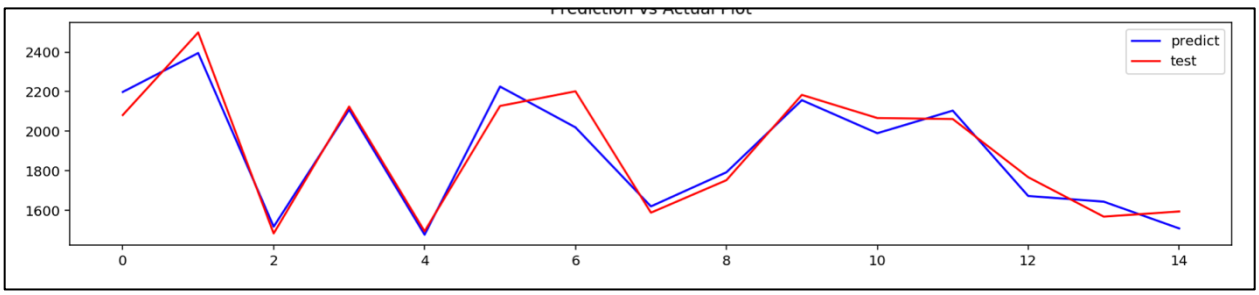
```
Mean Absolute Error: 69.40731341290339
Mean Squared Error: 6796.178504100094
Root Mean Squared Error: 82.43893803355363
```

The error metrics for the residential sector in Minnesota has a relatively low error, so it is promising that the linear model here provides a good representation of the data.

Actual vs Prediction:




Actual vs. Predict result scatter plot
Predicted
Actual
2400
2200
2000
1800
1600
1600
1800
2000
2200
2400


predict
test
2400
2200
2000
1800
1600
0
2
4
6
8
10
12
14

In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted. In all three graphs, we can tell that the model is a good representation of the relationship between the features that we select and the retail sales of electricity that we want to predict. The trend in both graphs are similar, and the values are quite close to each other. This indicates that our linear model is predicting the data well.

4. Missouri

4.1 Commercial Sector for Missouri

4.1.1 VIF

```
Commercial_Retail Price          6.678167
CLDD                            42.688881
TAVG                           490.770613
AWND                             3.362815
HTDD                           304.817924
area                        149925.510047
population                       3.618034
solar-generation                 7.161409
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD and CLDD have large VIF, which means that the variables are multicollinear, as a

result, we try to remove area and TAVG to reduce the multicollinearity from the dataset.

```
const                          148530.584647
Commercial_Retail Price             6.671957
CLDD                                7.173197
AWND                                3.360848
HTDD                                5.219317
population                          3.564716
solar-generation                    7.083281
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

4.1.2 Regression on the Commercial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	Commercial_Usage	R-squared:	0.934
Model:	OLS	Adj. R-squared:	0.921
Method:	Least Squares	F-statistic:	73.44
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	1.23e-24
Time:	14:44:50	Log-Likelihood:	-311.88
No. Observations:	57	AIC:	643.8
Df Residuals:	47	BIC:	664.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2321.9713	90.352	25.699	0.000	2140.206	2503.737
Commercial_Retail Price	-1.1747	98.923	-0.012	0.991	-200.182	197.833
CLDD	705.8229	73.223	9.639	0.000	558.518	853.128
AWND	-239.1151	69.060	-3.462	0.001	-378.045	-100.185
HTDD	407.0719	85.798	4.745	0.000	234.469	579.675
population	-61.8517	59.459	-1.040	0.304	-181.468	57.765
solar-generation	78.0088	127.088	0.614	0.542	-177.659	333.677
Summer	3.5476	50.482	0.070	0.944	-98.009	105.105
Fall	12.1563	33.269	0.365	0.716	-54.773	79.085
Winter	26.4937	44.426	0.596	0.554	-62.881	115.868

Omnibus:	1.868	Durbin-Watson:	1.644
Prob(Omnibus):	0.393	Jarque-Bera (JB):	1.333
Skew:	0.370	Prob(JB):	0.514
Kurtosis:	3.117	Cond. No.	29.9

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.934, and the adjusted R-squared is 0.921, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 93.4% of the overall result.

Feature coefficients and Significance

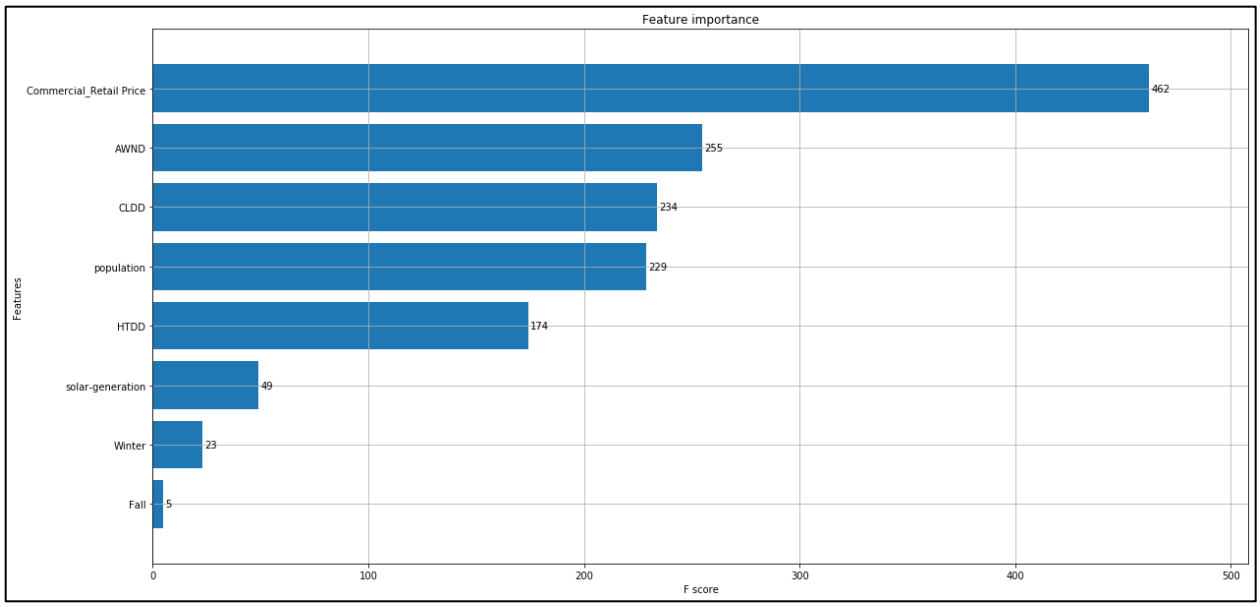
From the above table, we can say that CLDD, AWND and HTDD are significant features in the model. It means that for each increase in cooling degree days, the electricity consumption is expected to increase by 705.8299 mkWh.

MAE, MSE, RMSE:

```
Performance Evaluation
Mean Absolute Error: 70.62459736194344
Mean Squared Error: 7838.9722282538405
Root Mean Squared Error: 88.53797054514995
```

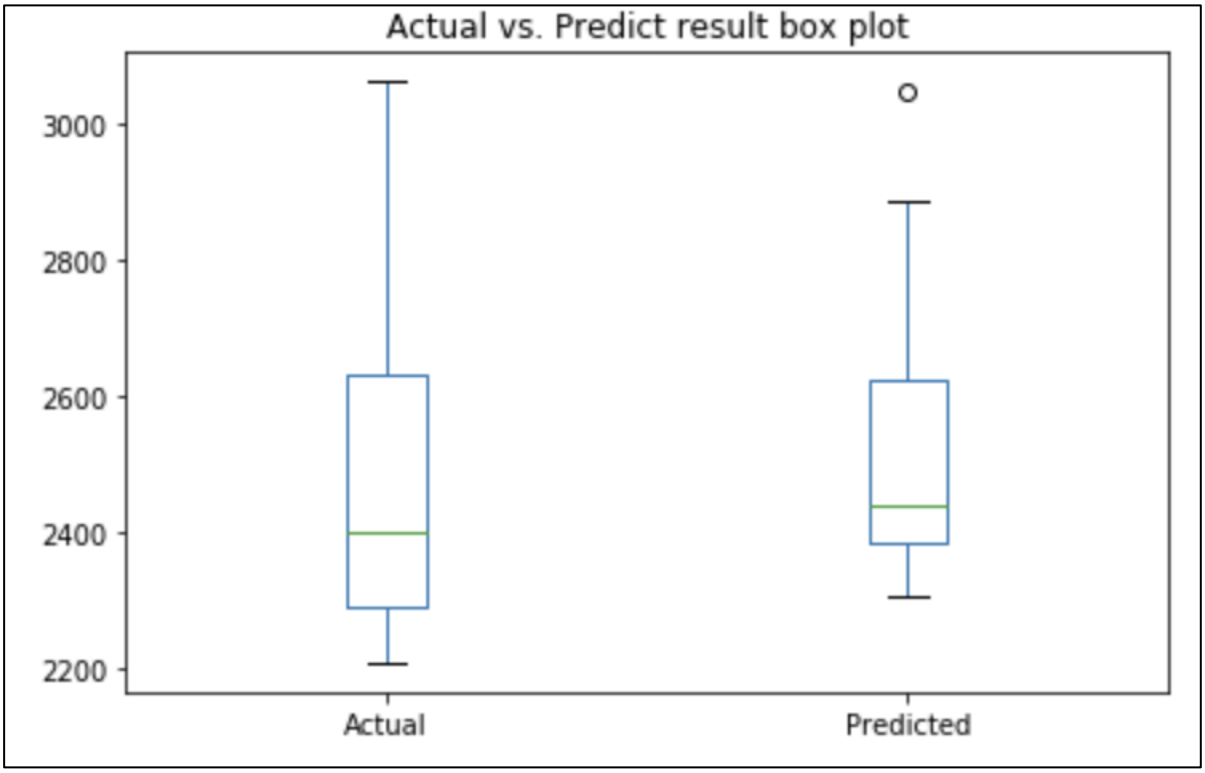
The error metrics for the commercial sector in Minnesota has a relatively low error, so it is promising that the linear model here provides a good representation of the data.

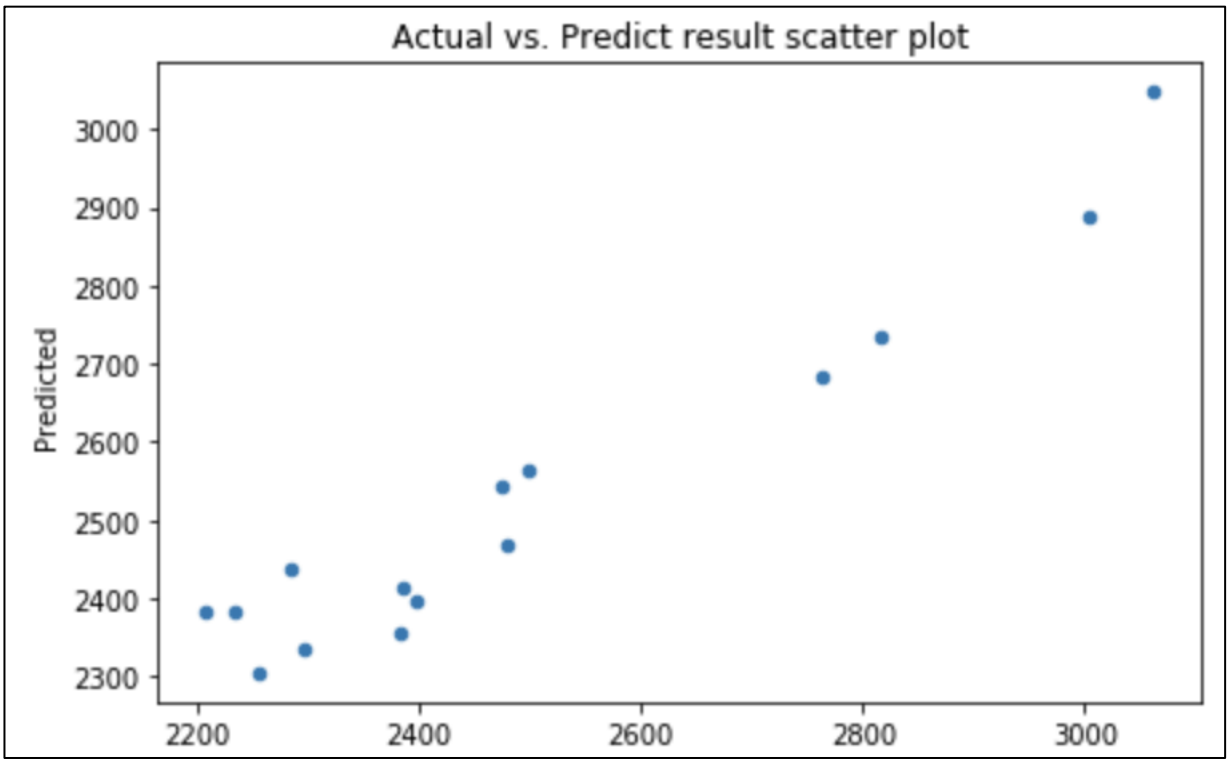
Feature Importance:

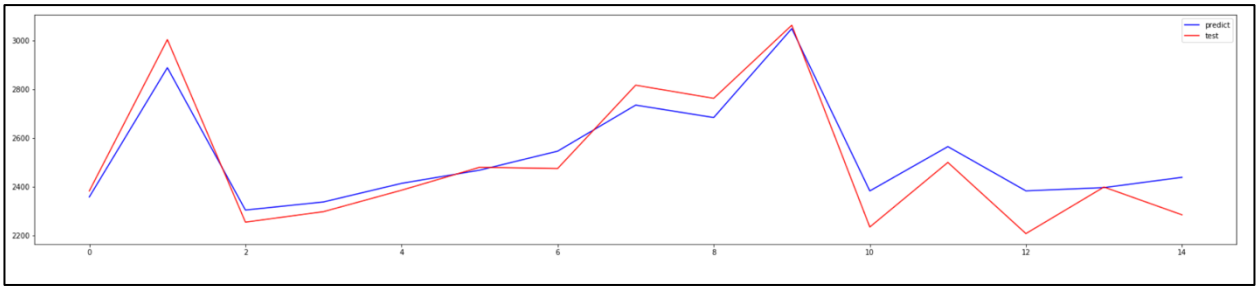


In the commercial sector, the most important feature is the price. Followed by the average wind days, cooling degree days and population. So when we want to understand the change of electricity consumption, these are features that we can look into.

Actual vs Predicted:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is a good representation of the relationship between the features that we select and the retail sales of electricity that we want to predict. The trend in both graphs are similar, in some places, they even fit almost the same. However, the issue with linear model is that it fails for extreme values that are outliers in the dataset.

4.2 Industrial Sector for Missouri

4.2.1 VIF

```
Industrial_Retail Price          4.630918
CLDD                            43.407788
TAVG                           490.566590
AWND                             3.385964
HTDD                           305.311183
area                        162068.541642
population                       4.078498
solar-generation                 7.250089
dtype: float64
```

When we put all the variables into the model, we can see that area, TAVG, HTDD and CLDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove area and TAVG to reduce the multicollinearity from the dataset.

```
const                       160757.592674
Industrial_Retail Price          4.628536
CLDD                             6.821266
AWND                             3.383331
HTDD                             5.077587
population                       4.020075
solar-generation                 7.171363
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

4.2.2 Regression on the industrial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Industrial_Usage   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.884
Method:                 Least Squares   F-statistic:                     48.51
Date:                Tue, 21 Jul 2020   Prob (F-statistic):           8.55e-21
Time:                        14:45:09   Log-Likelihood:                -316.05
No. Observations:                  57   AIC:                             652.1
Df Residuals:                      47   BIC:                             672.5
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                       1610.5247     89.009     18.094      0.000    1431.461    1789.589
Industrial_Retail Price     -196.7144     78.617     -2.502      0.016    -354.872     -38.556
CLDD                         184.1903     80.141      2.298      0.026      22.967     345.414
AWND                        -183.9077     72.927     -2.522      0.015    -330.617     -37.199
HTDD                          23.4209     86.974      0.269      0.789    -151.549     198.391
population                  -469.5556     71.394     -6.577      0.000    -613.182    -325.929
solar-generation            -148.9861    141.855     -1.050      0.299    -434.362     136.390
Summer                        23.5142     49.566      0.474      0.637     -76.201     123.229
Fall                           9.1085     34.480      0.264      0.793     -60.255      78.472
Winter                       -60.1240     47.854     -1.256      0.215    -156.393      36.145
==============================================================================
Omnibus:                        2.855   Durbin-Watson:                   2.429
Prob(Omnibus):                  0.240   Jarque-Bera (JB):                2.658
Skew:                          -0.455   Prob(JB):                        0.265
Kurtosis:                       2.461   Cond. No.                         30.7
==============================================================================
```

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.903, and the adjusted R-squared is 0.884, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 90.3% of the overall result.

Feature coefficients and Significance

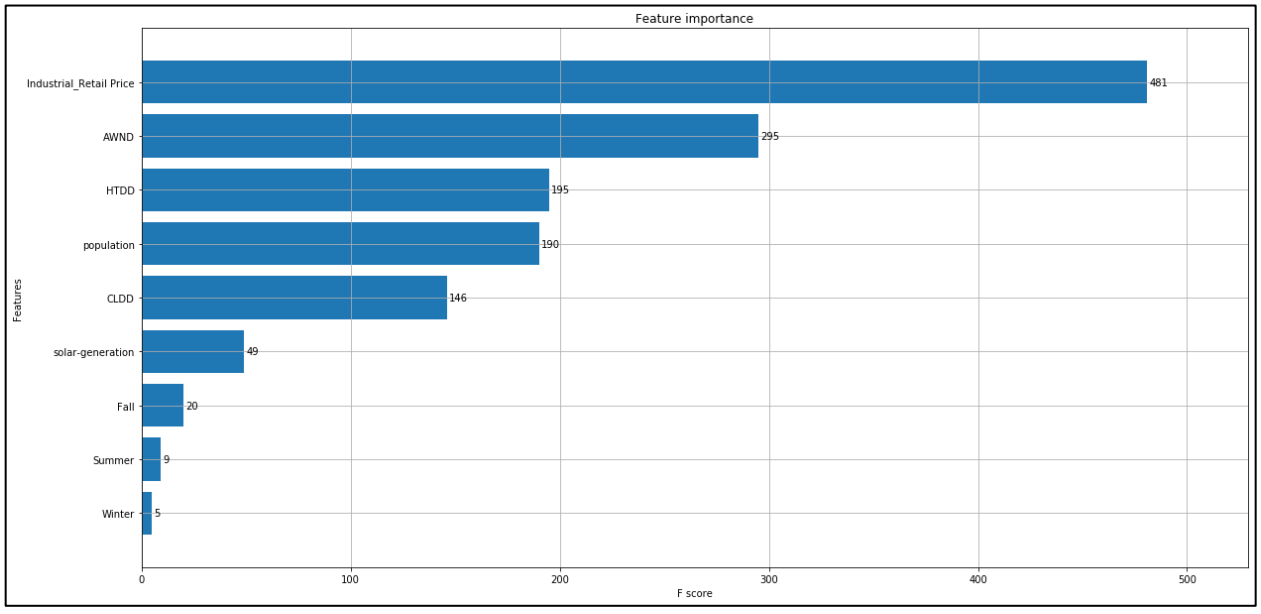
From the above table, we can say that price, CLDD, AWND and population is a significant variable of our model. The interpretation is that with each increase in price of electricity, there will be a corresponding decrease of 196.7144 mkWh in electricity consumption.

MAE, MSE, RMSE:

```
Performance Evaluation
Mean Absolute Error: 62.79854597894029
Mean Squared Error: 5147.734695108286
Root Mean Squared Error: 71.74771560898846
```

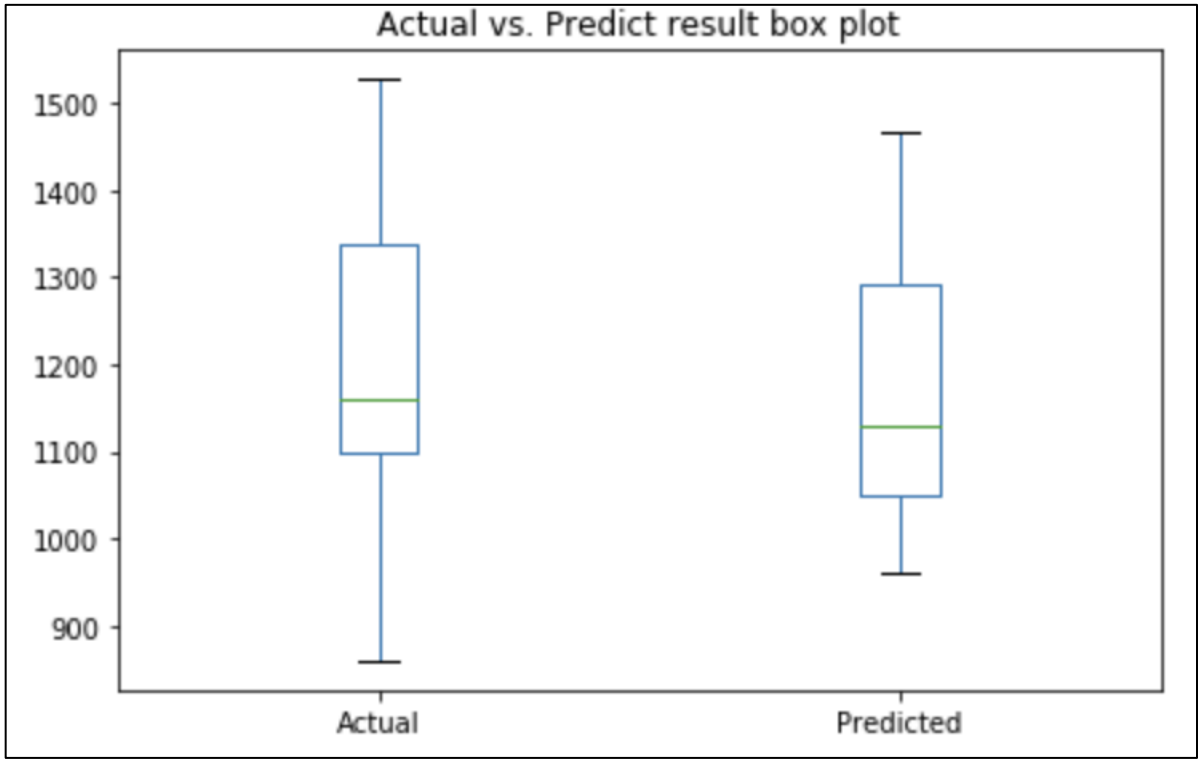
The error metrics for the commercial sector in Minnesota has a relatively low error, so it is promising that the linear model here provides a good representation of the data.

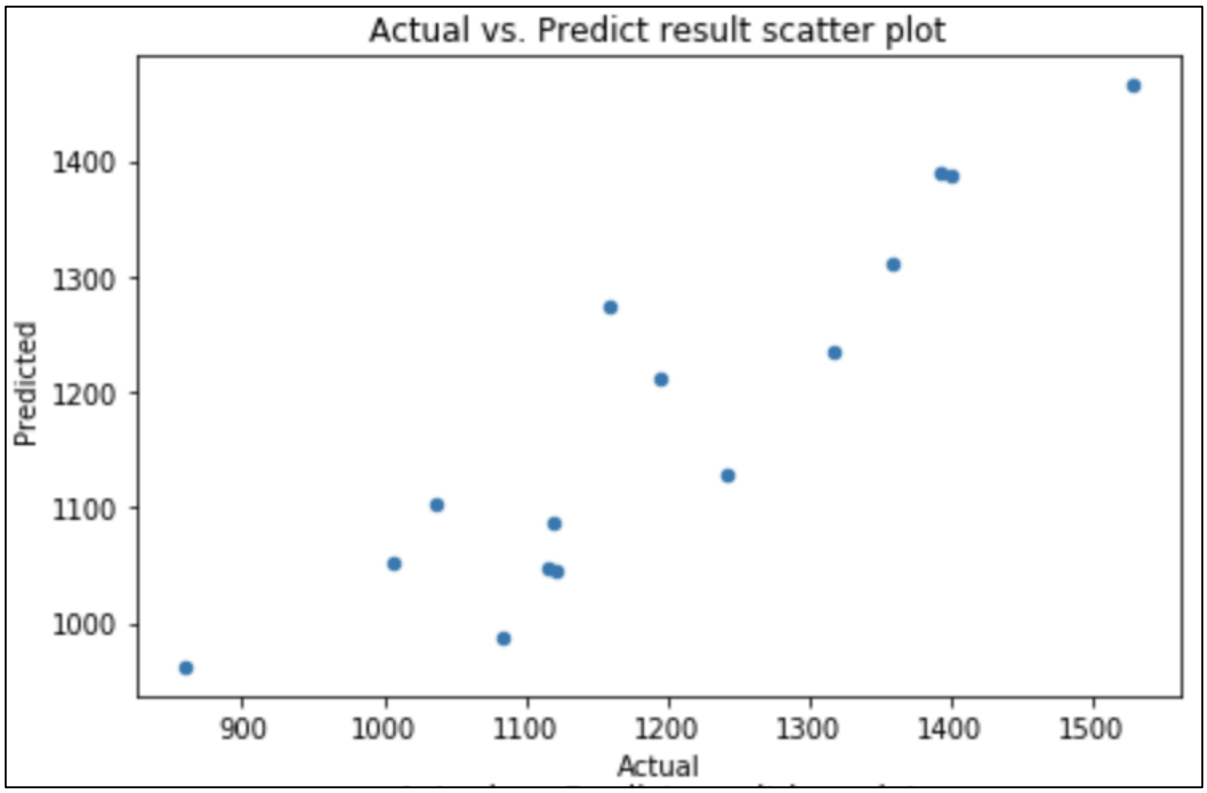
Feature importance:

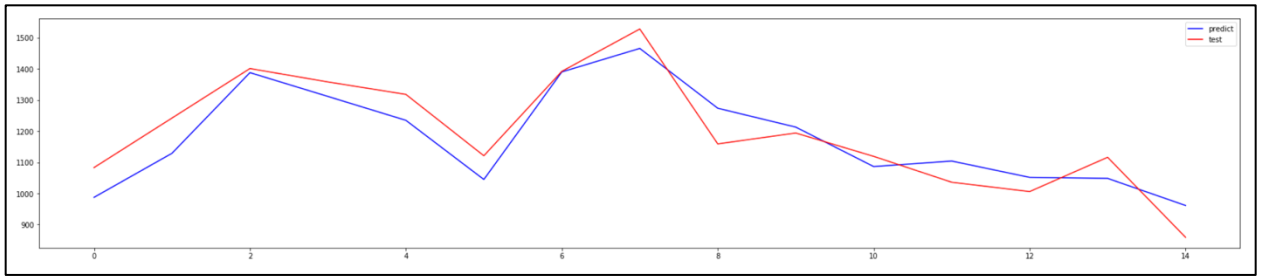


In the industrial sector, the price of electricity is the dominating feature that influence the consumption of electricity.

Actual vs Predicted







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is a good representation of the relationship between the features that we select and the retail sales of electricity that we want to predict. The trend in both graphs are similar, in some places, they even fit almost the same. However, the issue with linear model is that it fails for extreme values that are outliers in the dataset.

4.3 Residential Sector for Missouri

4.3.1 VIF

Residential_Retail Price	7.645370
CLDD	40.579353
TAVG	493.584791
AWND	3.360685
HTDD	304.826627
area	149464.082471
population	3.615929
solar-generation	7.198433
dtype: float64	

When we put all the variables into the model, we can see that area, TAVG, HTDD and CLDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove area and TAVG to reduce the multicollinearity from the dataset.

const	148246.831736
Residential_Retail Price	7.594712
CLDD	5.257238
AWND	3.357788
HTDD	7.263253
population	3.559169
solar-generation	7.110106
dtype: float64	

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

4.3.2 Regression on the Residential Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Residential_Usage   R-squared:                       0.950
Model:                            OLS   Adj. R-squared:                  0.941
Method:                 Least Squares   F-statistic:                     99.46
Date:                Tue, 21 Jul 2020   Prob (F-statistic):           1.58e-27
Time:                        14:45:19   Log-Likelihood:                -360.04
No. Observations:                  57   AIC:                             740.1
Df Residuals:                      47   BIC:                             760.5
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                       2209.9584    230.889      9.572      0.000    1745.469    2674.448
Residential_Retail Price    -717.9695    228.951     -3.136      0.003   -1178.560    -257.379
CLDD                        2031.5282    165.112     12.304      0.000    1699.366    2363.690
AWND                        -454.6870    157.768     -2.882      0.006    -772.075    -137.299
HTDD                        1782.2278    217.162      8.207      0.000    1345.353    2219.102
population                   -58.2093    137.709     -0.423      0.674    -335.244     218.825
solar-generation              82.6740    293.574      0.282      0.779    -507.922     673.270
Summer                       228.5451    109.100      2.095      0.042       9.065     448.025
Fall                         -82.7330     76.212     -1.086      0.283    -236.051      70.585
Winter                       217.5755    104.099      2.090      0.042       8.156     426.995
==============================================================================
Omnibus:                        4.501   Durbin-Watson:                   1.711
Prob(Omnibus):                  0.105   Jarque-Bera (JB):                3.490
Skew:                           0.549   Prob(JB):                        0.175
Kurtosis:                       3.514   Cond. No.                         30.7
==============================================================================
```

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.95, and the adjusted R-squared is 0.941, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 95% of the overall result.

Feature coefficients and Significance

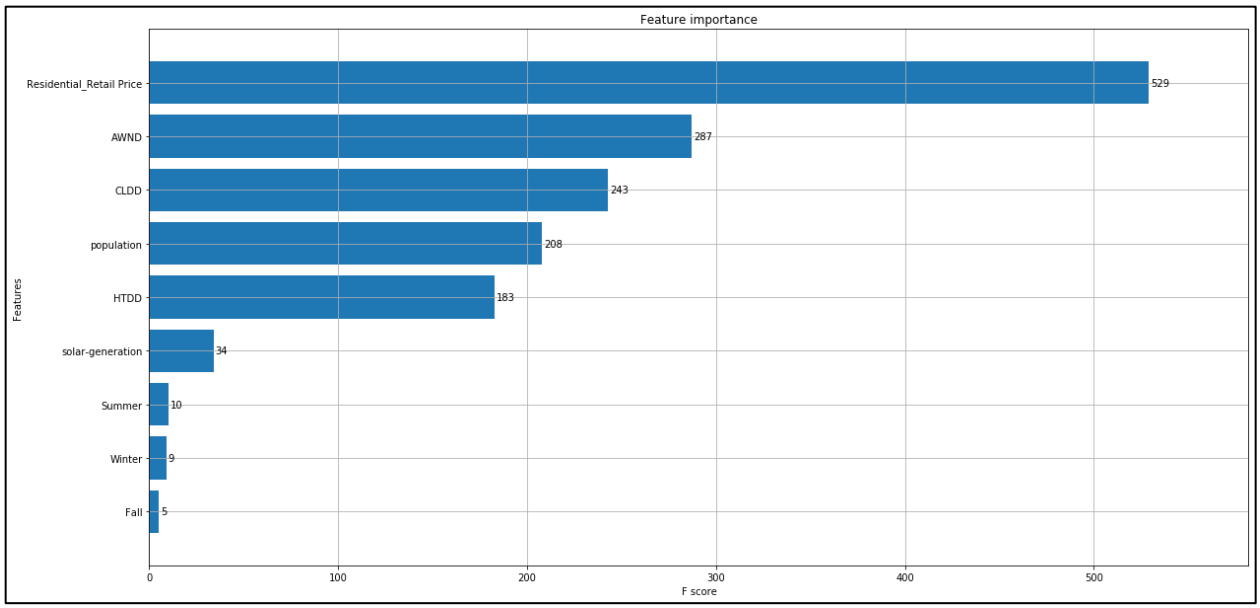
From the above table, we can say that price, CLDD, AWND and HTDD are significant features in the model. It means that for each increase in retail price of electricity, the electricity consumption is expected to decrease by 717.9695 mkWh.

MAE, MSE, RMSE:

```
Performance Evaluation
Mean Absolute Error: 147.7776583599763
Mean Squared Error: 27701.4802834782
Root Mean Squared Error: 166.437616792233
```

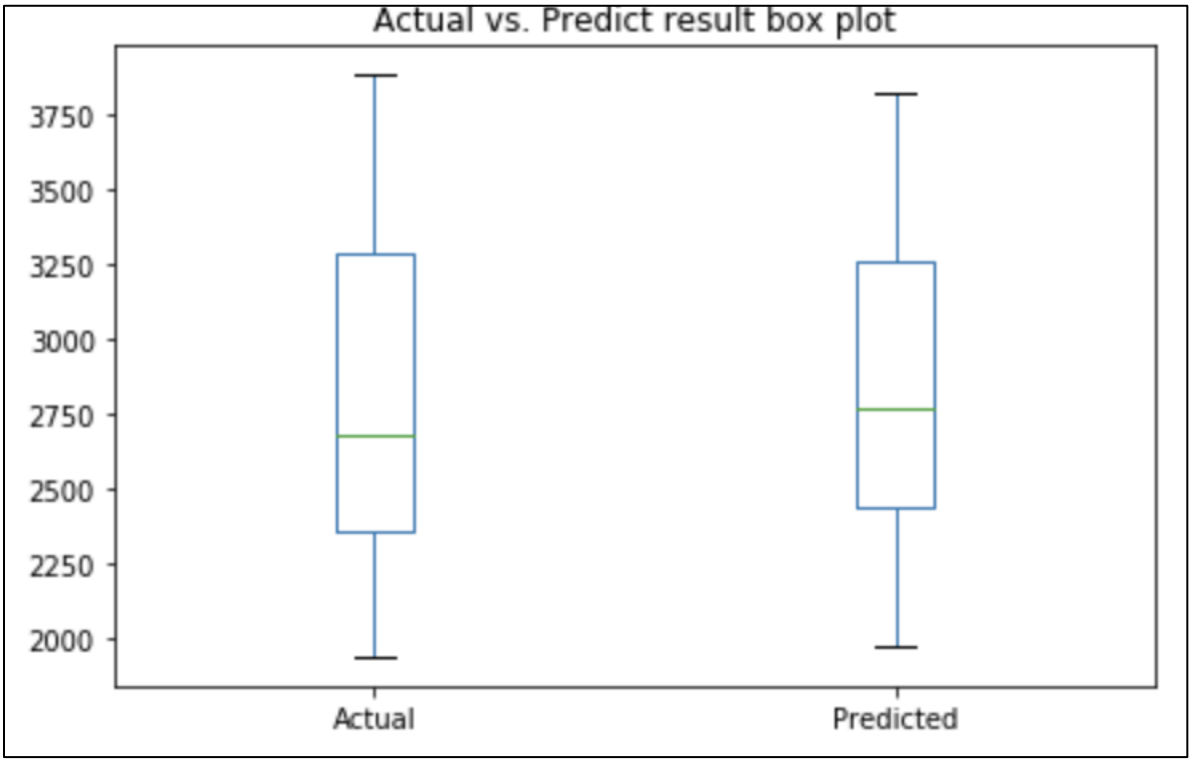
The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

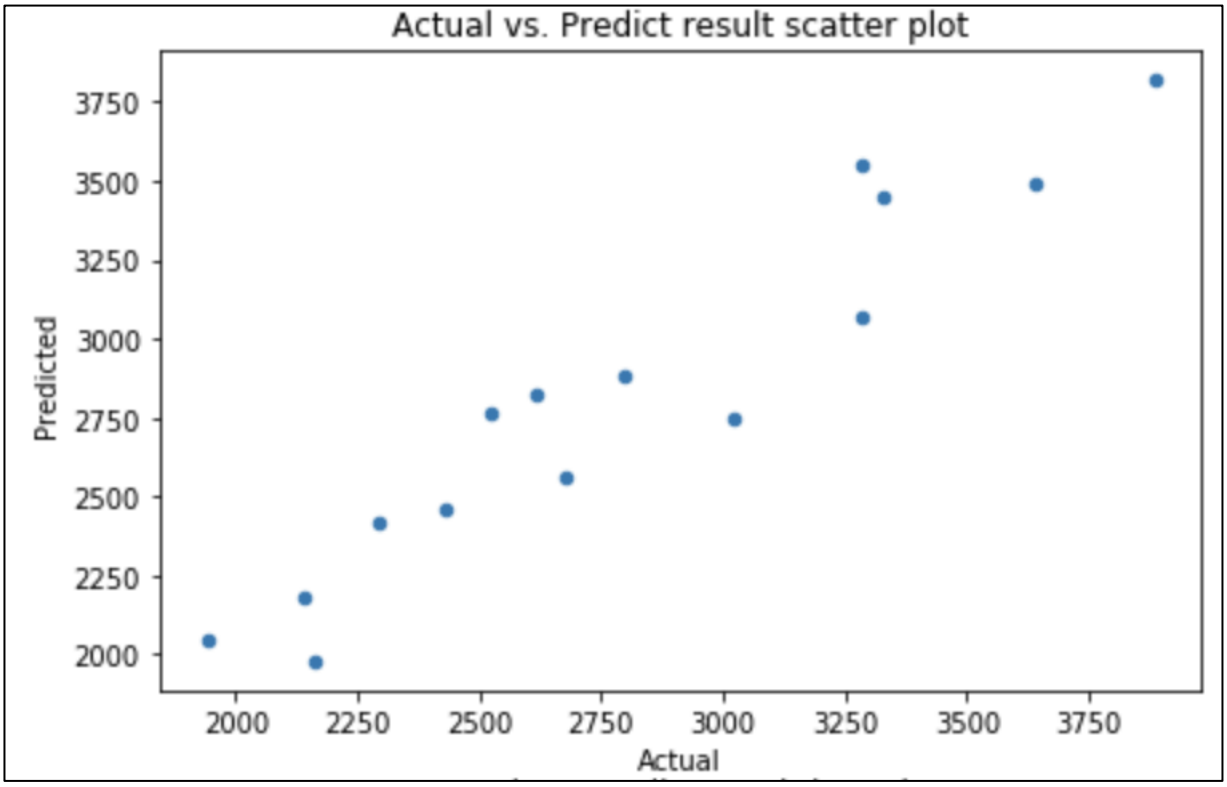
Feature Importance:

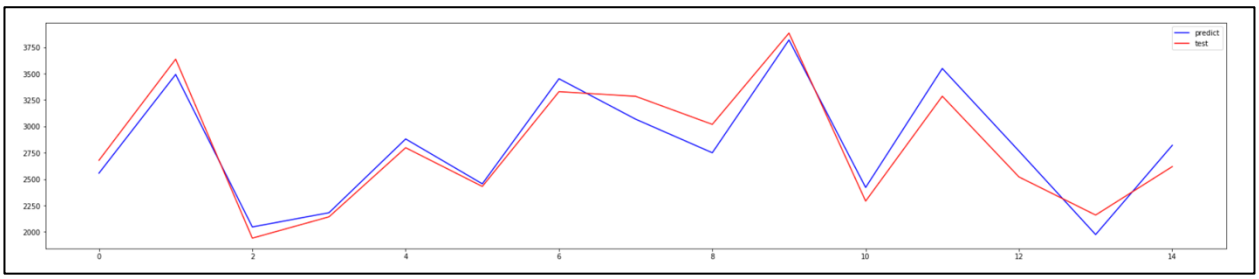


In the residential sector, the most important feature is the price of electricity. Followed by weather related features, average wind speed, cooling degree days and heating degree days. Then the population also influences the consumption of electricity.

Actual vs Predicted:




Actual vs. Predict result scatter plot
Predicted
Actual
2000
2250
2500
2750
3000
3250
3500
3750


predict
test

In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is a representation of the relationship between the features that we select and the retail sales of electricity that we want to predict. The trend in both graphs are similar, but the model fails to predict extreme values.

5 Texas

5.1 Commercial Sector for Texas

5.1.1 VIF

```
Commercial_Retail Price        1.179016
CLDD                         399.832182
TAVG                        1362.556995
AWND                           1.377606
HTDD                         364.674646
area                       44334.432347
population                     5.986158
solar-generation               7.379153
dtype: float64
```

When we put all the variables into the model, we can see that area, HTDD, TAVG and CLDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove area to reduce the multicollinearity from the dataset.

```
const                          9881.245184
Commercial_Retail Price           1.159551
CLDD                              3.700600
AWND                              1.360703
HTDD                              3.015056
population                        5.964476
solar-generation                  7.363510
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

5.1.2 Regression on the Commercial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Commercial_Usage   R-squared:                       0.901
Model:                            OLS   Adj. R-squared:                  0.882
Method:                 Least Squares   F-statistic:                     47.30
Date:                Tue, 21 Jul 2020   Prob (F-statistic):           1.44e-20
Time:                        13:49:36   Log-Likelihood:                -431.21
No. Observations:                  57   AIC:                             882.4
Df Residuals:                      47   BIC:                             902.8
Df Model:                           9
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     9928.5878    478.340     20.756      0.000    8966.292    1.09e+04
Commercial_Retail Price   -249.7536    326.806     -0.764      0.449    -907.203     407.696
CLDD                      5036.7794    761.344      6.616      0.000    3505.153    6568.406
AWND                      -809.0743    365.101     -2.216      0.032   -1543.562     -74.586
HTDD                      1514.0228    635.137      2.384      0.021     236.293    2791.753
population                 206.6782    664.575      0.311      0.757   -1130.273    1543.630
solar-generation          -109.6859    855.964     -0.128      0.899   -1831.664    1612.292
Summer                    -213.2444    423.241     -0.504      0.617   -1064.695     638.207
Fall                       360.8044    310.635      1.162      0.251    -264.113     985.722
Winter                    -190.4835    374.098     -0.509      0.613    -943.071     562.104
==============================================================================
Omnibus:                        7.546   Durbin-Watson:                   2.217
Prob(Omnibus):                  0.023   Jarque-Bera (JB):                6.887
Skew:                           0.827   Prob(JB):                       0.0320
Kurtosis:                       3.406   Cond. No.                         28.5
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.901, and the adjusted R-squared is 0.882, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 90.1% of the overall result.

Feature coefficients and Significance

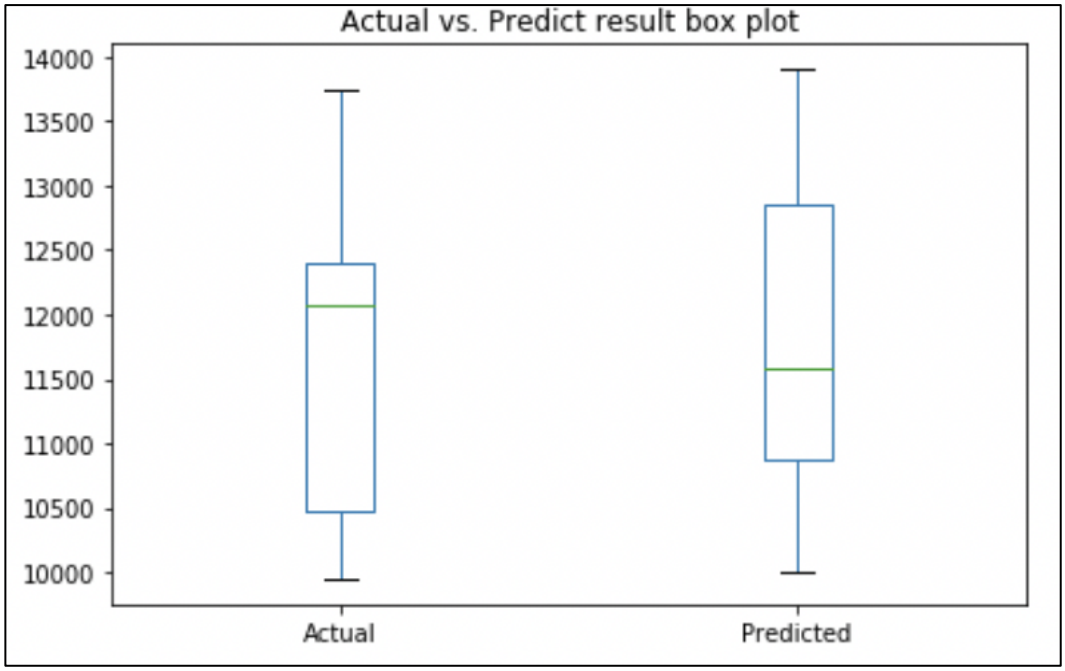
From the above table, we can say that price, CLDD, AWND and HTDD are significant features in the model. The logic is that with each increase in heating degree days, there will be a corresponding 1514.0228 mkWh increase in electricity consumption

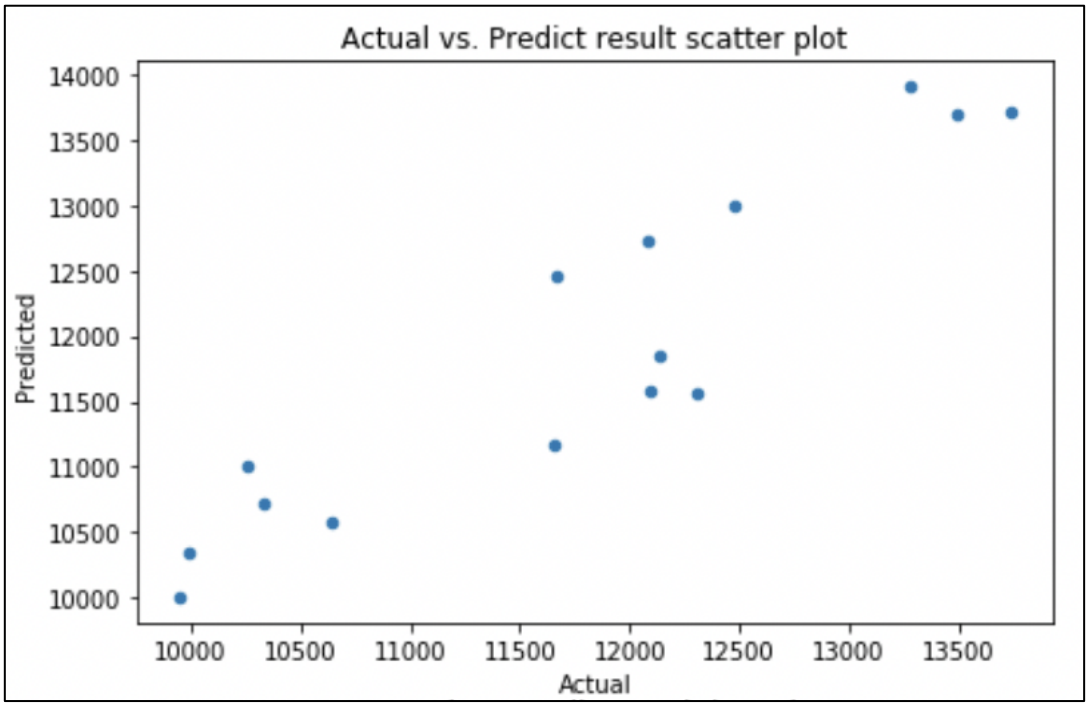
MAE, MSE, RMSE:

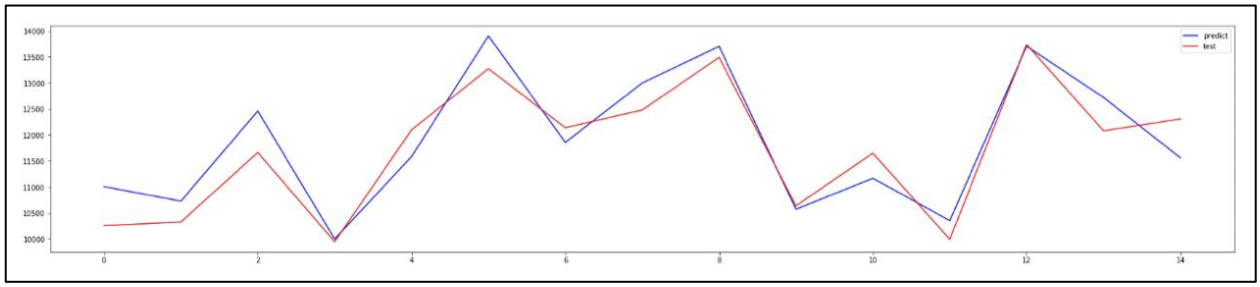
```
Performance Evaluation
Mean Absolute Error: 432.31528876282755
Mean Squared Error: 250126.39971848117
Root Mean Squared Error: 500.12638374563
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Predicted:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is making a good prediction of the test data. The trends in actual data and predicted value is quite close. The reason for the high error may be due to high electricity consumption in certain datapoints that makes the value of the error high.

5.2 Industrial Sector for Texas

5.2.1 VIF

Industrial_Retail Price	1.700063
CLDD	406.757273
TAVG	1370.023185
AWND	1.330015
HTDD	363.106745
area	47487.685123
population	6.719965
solar-generation	7.054668
dtype: float64	

When we put all the variables into the model, we can see that area, HTDD, TAVG and CLDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove area to reduce the multicollinearity from the dataset.

const	16129.711762
Residential_Retail Price	1.443779
CLDD	4.080290
AWND	1.320913
HTDD	3.072831
population	7.809881
solar-generation	9.179434
dtype: float64	

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

5.2.2 Regression on the Industrial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Industrial_Usage   R-squared:                       0.792
Model:                            OLS   Adj. R-squared:                  0.752
Method:                 Least Squares   F-statistic:                     19.86
Date:                Tue, 21 Jul 2020   Prob (F-statistic):           3.28e-13
Time:                        13:49:41   Log-Likelihood:                -409.51
No. Observations:                  57   AIC:                             839.0
Df Residuals:                      47   BIC:                             859.5
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                     8730.1370    307.924     28.352      0.000    8110.674    9349.599
Industrial_Retail Price   -440.4946    353.531     -1.246      0.219   -1151.708     270.719
CLDD                      2011.6048    523.021      3.846      0.000     959.423    3063.787
AWND                      -164.0428    249.223     -0.658      0.514    -665.415     337.329
HTDD                       722.9193    447.371      1.616      0.113    -177.075    1622.913
population                2429.8448    498.379      4.875      0.000    1427.236    3432.454
solar-generation         -1776.4322    592.176     -3.000      0.004   -2967.737    -585.127
Summer                    -312.8383    286.217     -1.093      0.280    -888.632     262.955
Fall                      -469.9957    218.403     -2.152      0.037    -909.365     -30.626
Winter                    -783.1030    253.235     -3.092      0.003   -1292.546    -273.660
==============================================================================
Omnibus:                        2.372   Durbin-Watson:                   2.261
Prob(Omnibus):                  0.305   Jarque-Bera (JB):                1.721
Skew:                           0.416   Prob(JB):                        0.423
Kurtosis:                       3.179   Cond. No.                         26.7
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.792, and the adjusted R-squared is 0.752, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 75.2% of the overall result.

Feature coefficients and Significance

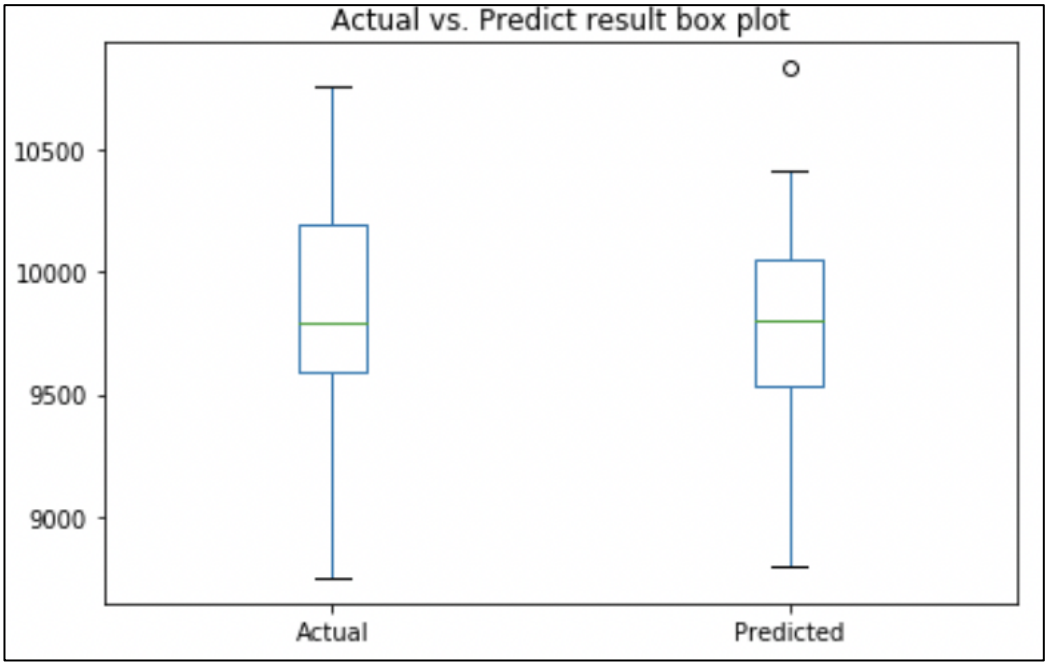
From the above table, we can say that CLDD, population, solar-generation and Winter are significant features in the model. The logic is that with each increase in electricity generated by distributed solar photovoltaic, there will be a corresponding 1776.4322 mkWh decrease in electricity consumption.

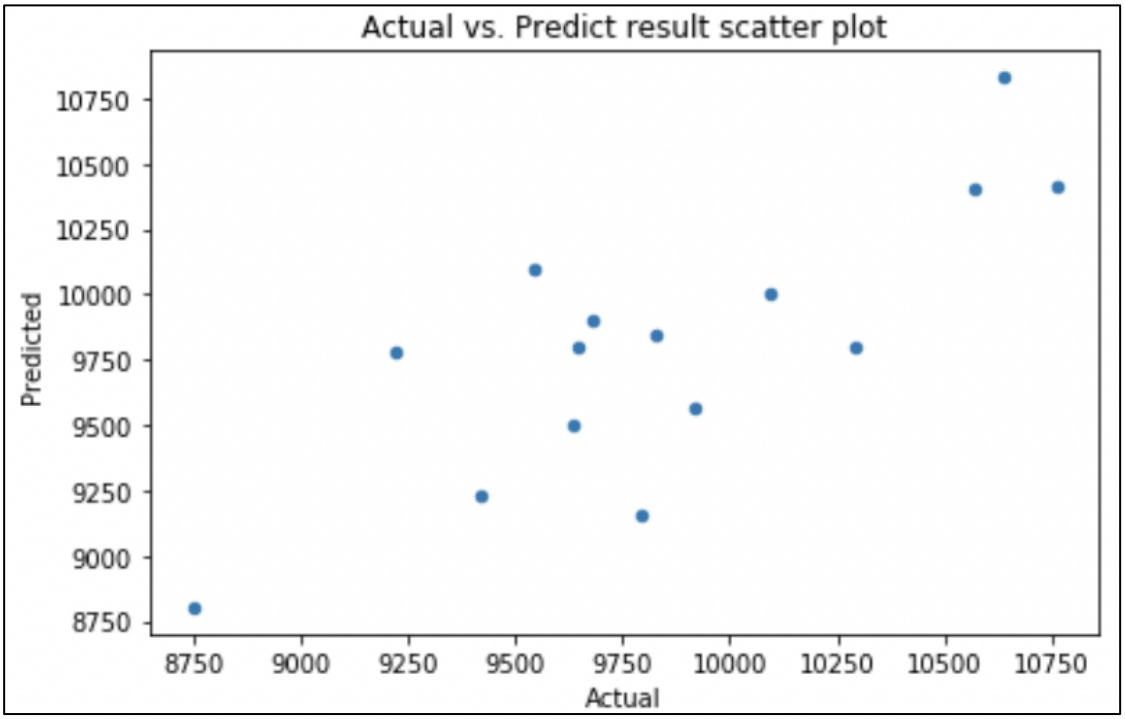
MAE, MSE, RMSE:

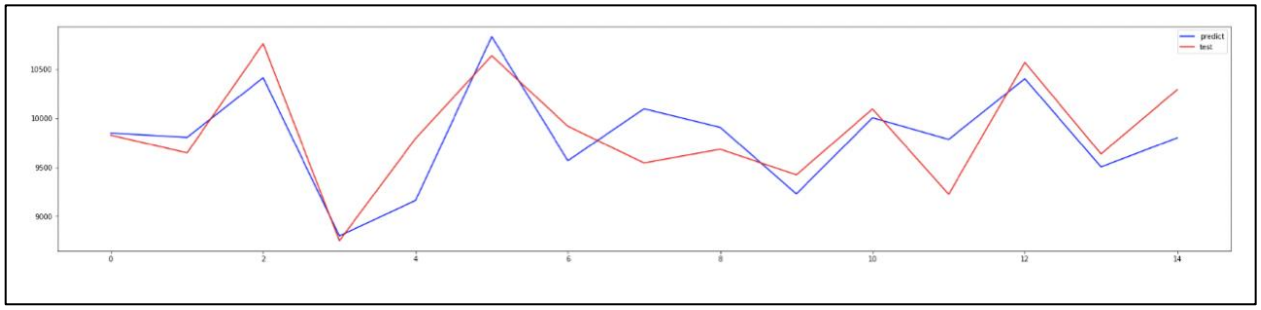
```
Performance Evaluation
Mean Absolute Error: 277.47137197652376
Mean Squared Error: 113896.79234111765
Root Mean Squared Error: 337.4859883626543
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Predicted:


Actual vs. Predict result box plot
10500
10000
9500
9000
Actual
Predicted


Actual vs. Predict result scatter plot
Predicted
10750
10500
10250
10000
9750
9500
9250
9000
8750
8750
9000
9250
9500
9750
10000
10250
10500
10750
Actual



In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is making a good prediction of the test data. The trends in actual data and predicted value is quite close. The reason for the high error may be due to high electricity consumption in certain datapoints that makes the value of the error high. There are outliers in the dataset that are not representable by our linear model.

5.3 Residential Sector for Texas

5.3.1 VIF

```
Residential_Retail Price        1.444449
CLDD                          393.525225
TAVG                         1340.684224
AWND                            1.345681
HTDD                          358.230068
area                        47873.783980
population                      7.826321
solar-generation                9.183411
dtype: float64
```

When we put all the variables into the model, we can see that area, HTDD, TAVG and CLDD have large VIF, which means that the variables are multicollinear, as a result, we try to remove area to reduce the multicollinearity from the dataset.

```
const                       11512.540052
Industrial_Retail Price         1.662883
CLDD                            4.364797
AWND                            1.301967
HTDD                            3.363044
population                      6.718157
solar-generation                7.053020
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

5.3.2 Regression on the Residential Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Residential_Usage   R-squared:                       0.934
Model:                            OLS   Adj. R-squared:                  0.921
Method:                 Least Squares   F-statistic:                     73.54
Date:                Tue, 21 Jul 2020   Prob (F-statistic):           1.20e-24
Time:                        13:49:47   Log-Likelihood:                -461.76
No. Observations:                  57   AIC:                             943.5
Df Residuals:                      47   BIC:                             964.0
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                       7132.7308    824.931      8.646      0.000    5473.183    8792.278
Residential_Retail Price -1648.7715    588.747     -2.800      0.007   -2833.177    -464.366
CLDD                        1.237e+04   1325.304      9.335      0.000    9705.176     1.5e+04
AWND                       -1456.2193    634.966     -2.293      0.026   -2733.607    -178.832
HTDD                        6684.1084   1087.015      6.149      0.000    4497.317    8870.900
population                 -2074.9914   1404.002     -1.478      0.146   -4899.479     749.496
solar-generation            3622.0491   1758.184      2.060      0.045      85.039    7159.059
Summer                      -432.1010    724.864     -0.596      0.554   -1890.339    1026.137
Fall                         855.6561    589.189      1.452      0.153    -329.639    2040.951
Winter                       806.6734    639.957      1.261      0.214    -480.753    2094.100
==============================================================================
Omnibus:                        0.703   Durbin-Watson:                   2.098
Prob(Omnibus):                  0.704   Jarque-Bera (JB):                0.645
Skew:                           0.247   Prob(JB):                        0.724
Kurtosis:                       2.836   Cond. No.                         33.5
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.934, and the adjusted R-squared is 0.921, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 93.4% of the overall result.

Feature coefficients and Significance

From the above table, we can say that price, CLDD, AWND, HTDD and solar generation are significant features in the model. The logic is that with each increase in heating degree days, there will be a corresponding 6684.1082 mkWh decrease in electricity consumption.

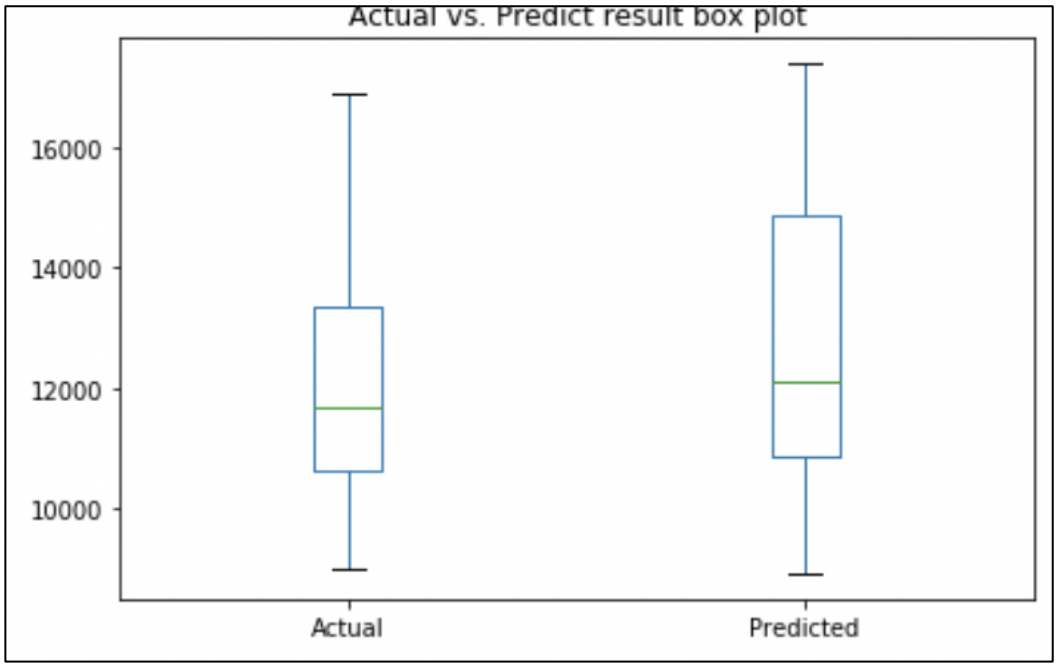
MAE, MSE, RMSE:

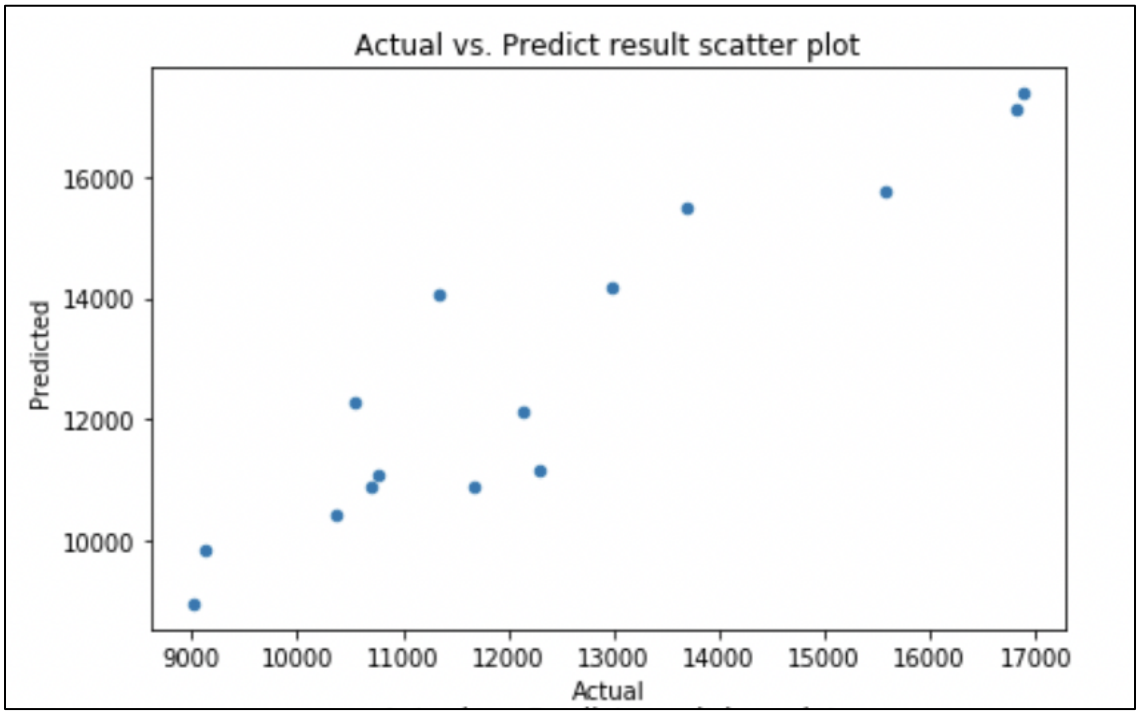
```
Performance Evaluation
Mean Absolute Error: 778.329963849823
Mean Squared Error: 1203762.0327639407
Root Mean Squared Error: 1097.1608964796096
```

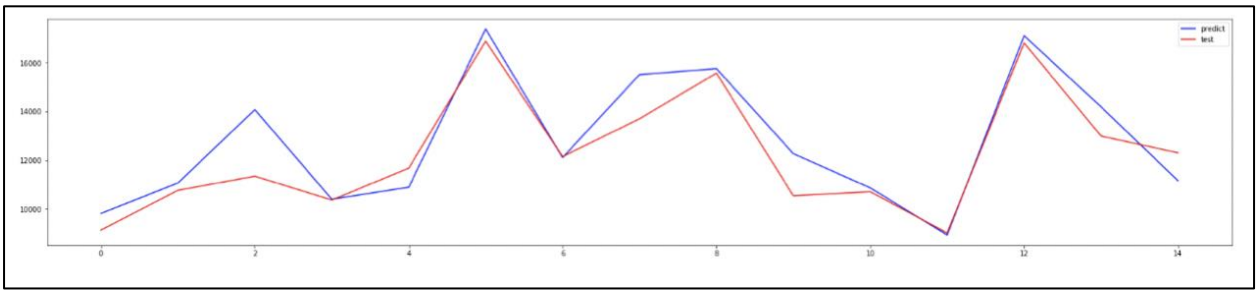
The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

In this model, the error metrics are high, which may be an indicator that our model is not representing the dataset well.

Actual vs Prediction:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is making a good prediction of the test data. The trends in actual data and predicted value is quite close. The reason for the high error may be due to high electricity consumption in certain datapoints that makes the value of the error high.

6 New York

6.1 Commercial Sector for New York

6.1.1 VIF

```
Commercial_Retail Price          2.220341
CLDD                             8.077153
TAVG                           334.845890
AWND                             5.886504
HTDD                           281.756066
area                        340562.822144
population                       6.570350
solar-generation                 6.872330
dtype: float64
```

When we put all the variables into the model, we can see that area, HTDD, TAVG have large VIF, which means that the variables are multicollinear, as a result, we try to remove area and TAVG to reduce the multicollinearity from the dataset.

```
const                       334523.976610
Commercial_Retail Price          2.192524
CLDD                             2.840211
AWND                             5.884754
HTDD                             6.067612
population                       6.569477
solar-generation                 6.866836
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variable.

6.1.2 Regression on the Commercial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	Commercial_Usage	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	28.41
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	4.13e-16
Time:	17:36:53	Log-Likelihood:	-380.87
No. Observations:	57	AIC:	781.7
Df Residuals:	47	BIC:	802.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5263.8421	378.357	13.912	0.000	4502.686	6024.998
Commercial_Retail Price	461.3991	186.094	2.479	0.017	87.026	835.772
CLDD	1604.7709	222.539	7.211	0.000	1157.081	2052.461
AWND	-225.1572	280.047	-0.804	0.425	-788.538	338.224
HTDD	604.7194	290.999	2.078	0.043	19.305	1190.134
population	280.2531	283.345	0.989	0.328	-289.764	850.270
solar-generation	339.0285	346.496	0.978	0.333	-358.032	1036.089
Summer	107.3759	142.153	0.755	0.454	-178.600	393.352
Fall	303.3227	124.563	2.435	0.019	52.734	553.912
Winter	321.8015	119.818	2.686	0.010	80.759	562.844

Omnibus:	0.184	Durbin-Watson:	2.110
Prob(Omnibus):	0.912	Jarque-Bera (JB):	0.266
Skew:	0.126	Prob(JB):	0.875
Kurtosis:	2.780	Cond. No.	32.2

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.845, and the adjusted R-squared is 0.815, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 84.5% of the overall result.

Feature coefficients and Significance

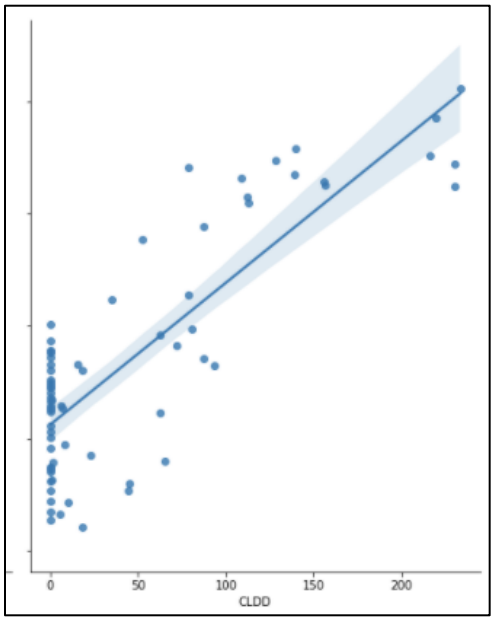
From the above table, we can say that price, CLDD, HTDD, Fall and Winter are significant features in the model. The logic is that since we choose Spring as the baseline season, when we are in Winter, the electricity consumption is expected to be 321.8015 mkWh higher than in Spring.

MAE, MSE, RMSE:

```
Performance Evaluation
Mean Absolute Error: 170.52748902617907
Mean Squared Error: 39899.5887284391
Root Mean Squared Error: 199.74881408518826
```

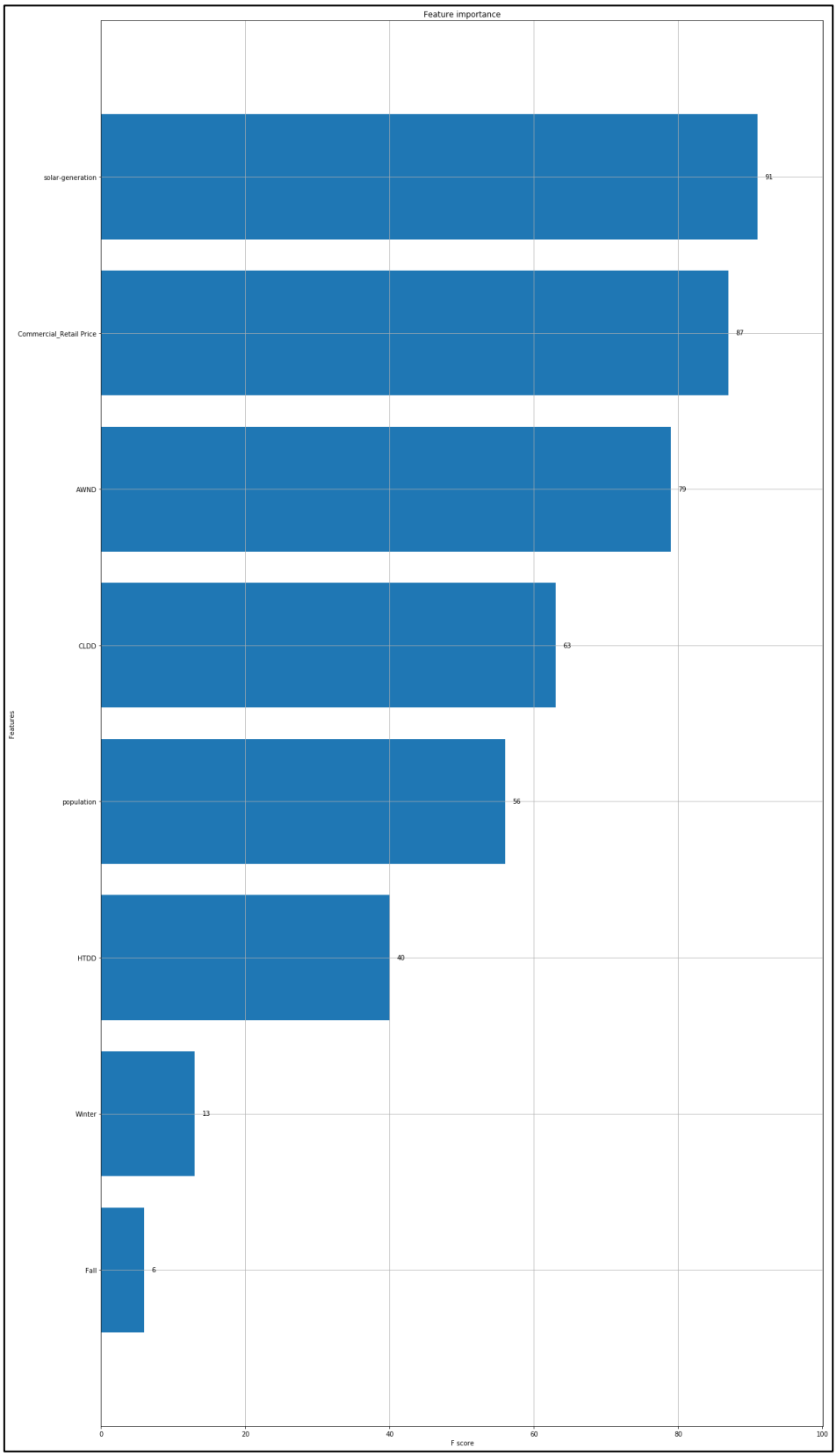
The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Correlation Analysis:



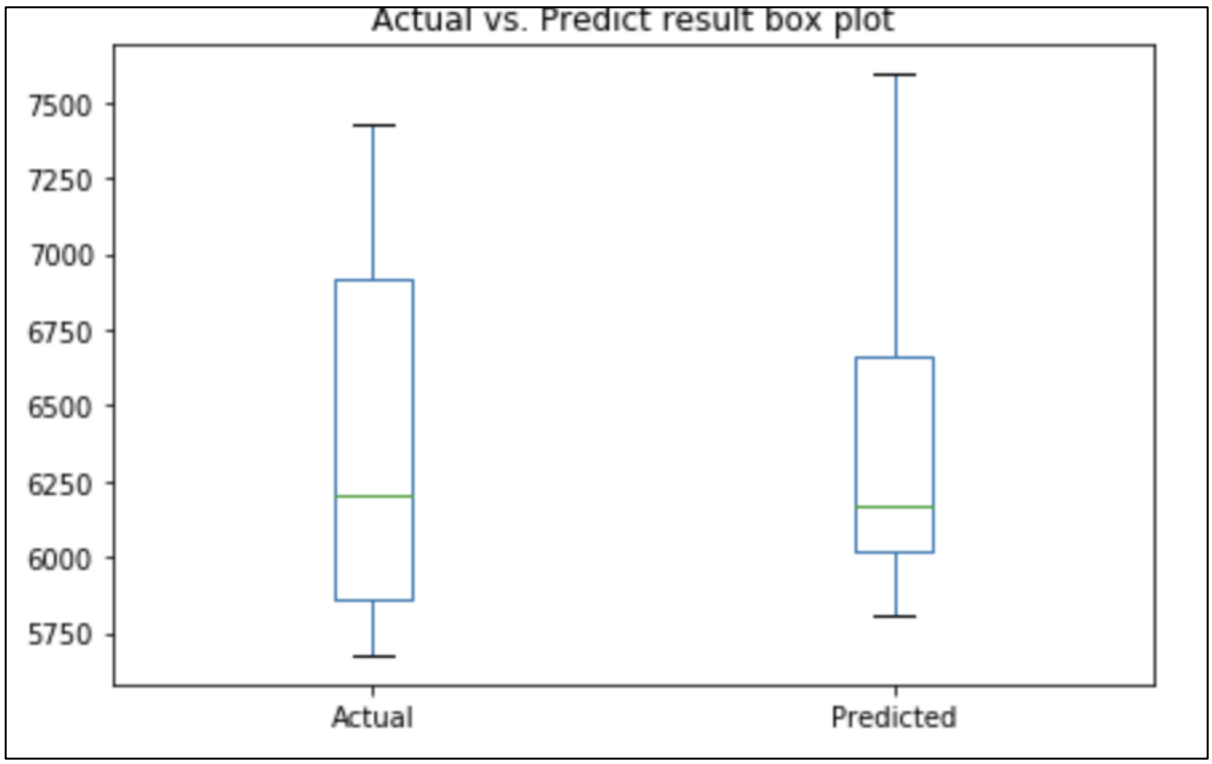
From the correlation figure, we can also find that cooling degree days are highly relevant to the electricity consumption in the commercial sector.

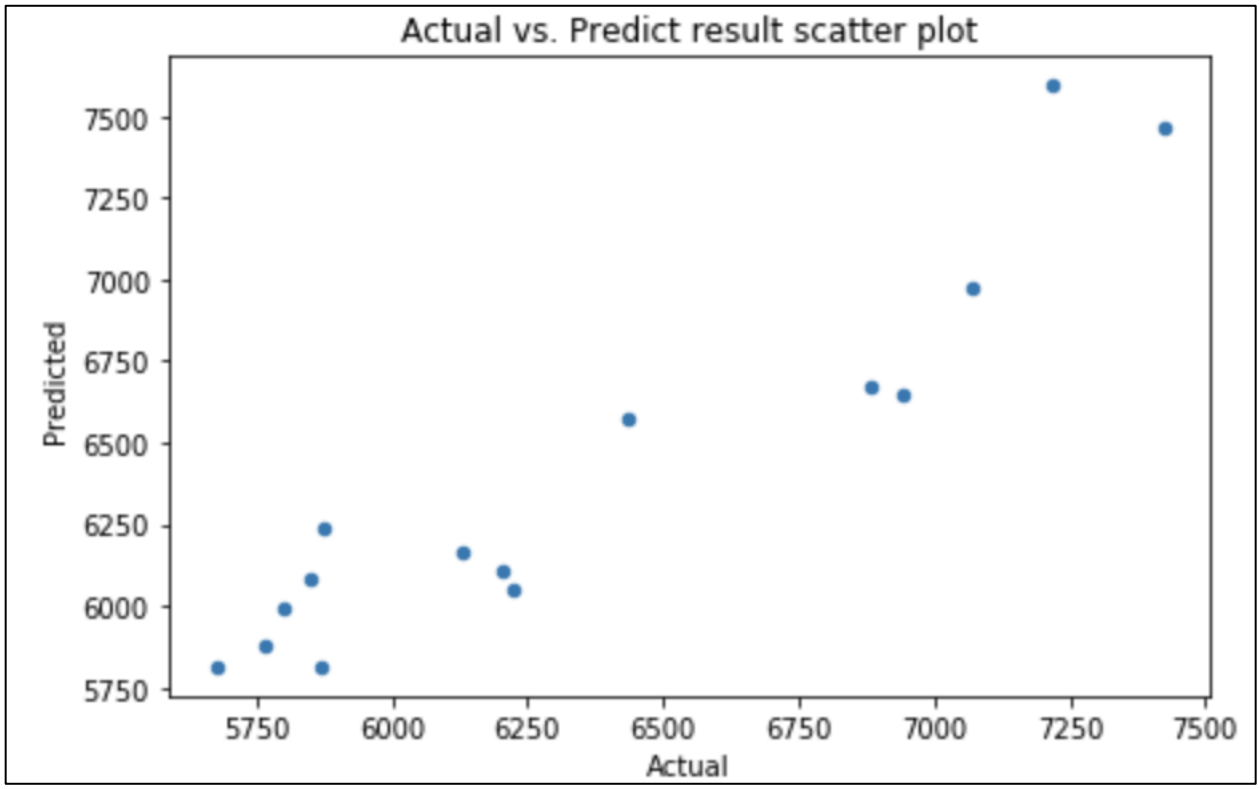
Feature Importance:

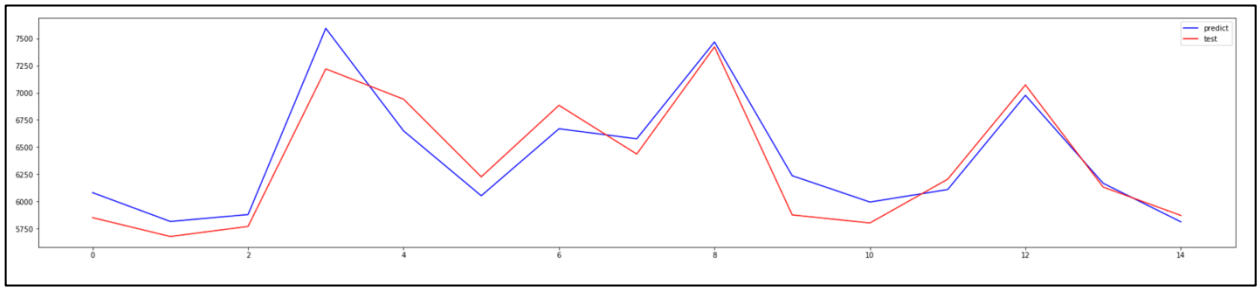


From the feature importance graph, we can see that electricity generated by distribute solar photovoltaic is an important feature in determining the electricity consumption in the commercial sector. Which indicates that the commercial sector may consume more distributed solar-generated electricity than retail sales of electricity.

Actual vs Predicted:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is making a good prediction of the test data. The trends in actual data and predicted value is quite close.

The reason for the high error may be due to high electricity consumption in certain datapoints that makes the value of the error high. There are outliers in the dataset that are not representable by our linear model

6.2 Industrial Sector for New York

6.2.1 VIF

```
Industrial_Retail Price          1.653804
CLDD                             7.887929
TAVG                           343.735017
AWND                             5.796430
HTDD                           284.704359
area                        365888.603851
population                       7.140433
solar-generation                 7.082131
dtype: float64
```

When we put all the variables into the model, we can see that area, HTDD, TAVG have large VIF, which means that the variables are multicollinear, as a result, we try to remove area and TAVG to reduce the multicollinearity from the dataset.

```
const                       363339.602378
Industrial_Retail Price          1.590853
CLDD                             2.676490
AWND                             5.786472
HTDD                             6.902401
population                       7.130933
solar-generation                 7.082066
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

6.2.2 Regression on the Industrial Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	Industrial_Usage	R-squared:	0.441
Model:	OLS	Adj. R-squared:	0.333
Method:	Least Squares	F-statistic:	4.112
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	0.000601
Time:	17:37:03	Log-Likelihood:	-337.04
No. Observations:	57	AIC:	694.1
Df Residuals:	47	BIC:	714.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1307.3803	175.238	7.461	0.000	954.848	1659.913
Industrial_Retail Price	27.9282	109.006	0.256	0.799	-191.364	247.220
CLDD	42.7311	103.747	0.412	0.682	-165.980	251.442
AWND	-2.7173	129.213	-0.021	0.983	-262.660	257.225
HTDD	81.0623	141.495	0.573	0.569	-203.588	365.713
population	71.8823	128.615	0.559	0.579	-186.858	330.622
solar-generation	-53.5392	159.751	-0.335	0.739	-374.916	267.838
Summer	198.0813	62.217	3.184	0.003	72.916	323.246
Fall	119.1396	55.761	2.137	0.038	6.962	231.317
Winter	14.3053	56.092	0.255	0.800	-98.538	127.149

Omnibus:	13.926	Durbin-Watson:	1.936
Prob(Omnibus):	0.001	Jarque-Bera (JB):	49.472
Skew:	0.290	Prob(JB):	1.81e-11
Kurtosis:	7.527	Cond. No.	31.5

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.441, and the adjusted R-squared is 0.333, there is a difference in the two metrics, indicating that some input variables are redundant. Furthermore, the 0.441 R-squared indicates that the model is explaining only 44.1% of the dataset, which means it may not be a good fit.

Feature coefficients and Significance

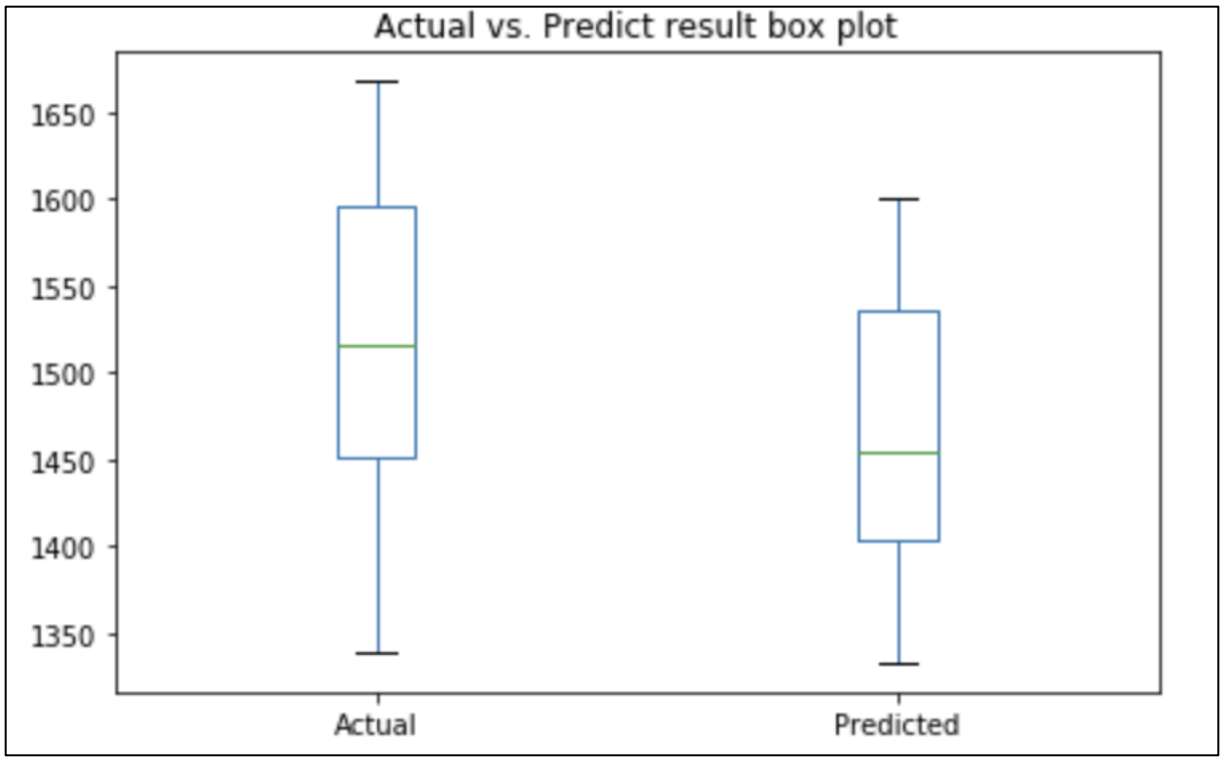
From the above table, we can say that Summer and Fall are significant features in the model. The logic is that since we used Spring as the baseline weather, when we enter Summer, the electricity consumption is expected to increase by 198.0813 mkWh.

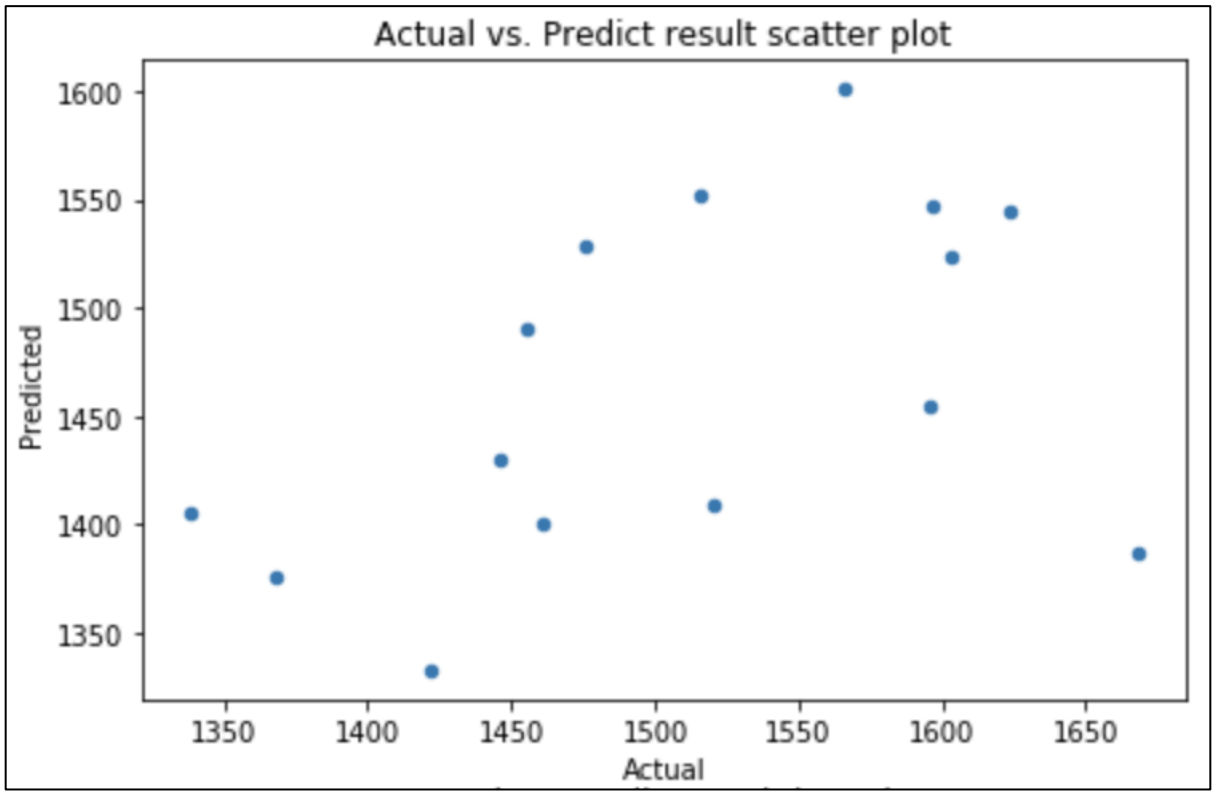
MAE, MSE, RMSE:

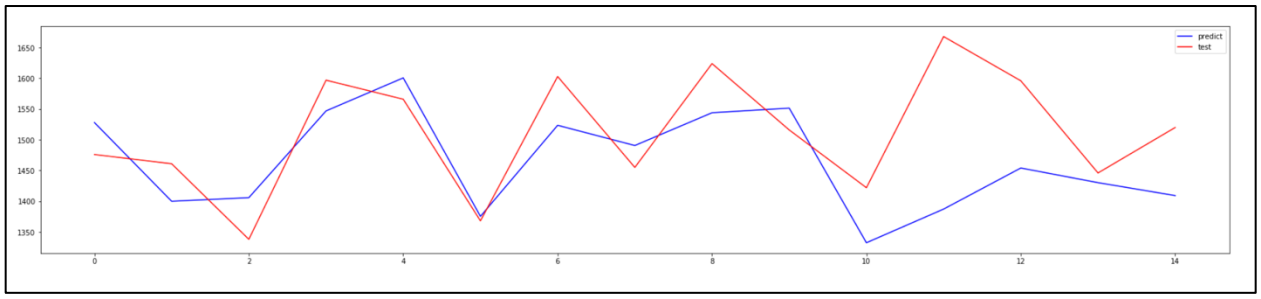
```
Performance Evaluation
Mean Absolute Error: 76.16799732497928
Mean Squared Error: 9972.669818797718
Root Mean Squared Error: 99.86325559883234
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. The error metrics here are comparatively low, so the actual predictions of our model may be good.

Actual vs Predicted:


Actual vs. Predict result box plot
1650
1600
1550
1500
1450
1400
1350
Actual
Predicted





In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is a poor representation of the dataset. There are huge differences in the predicted result and the actual data. We think this may indicate that the features we selected are not in a

linear relationship with the electricity consumption, so we will further explore their relationship in the ANN part.

6.3 Residential Sector for New York

6.3.1 VIF

```
Residential_Retail Price          1.332631
CLDD                              7.937355
TAVG                            348.488427
AWND                              5.472309
HTDD                            292.755591
area                         319979.391598
population                        6.090525
solar-generation                  6.864666
dtype: float64
```

When we put all the variables into the model, we can see that area, HTDD, TAVG have large VIF, which means that the variables are multicollinear, as a result, we try to remove area to reduce the multicollinearity from the dataset.

```
const                          313570.798723
Residential_Retail Price            1.264419
CLDD                                2.674296
AWND                                5.467884
HTDD                                5.990575
population                          6.090031
solar-generation                    6.858541
dtype: float64
```

After the removal, we can see that all the variables have a VIF score smaller than 10, so there is no multicollinearity among these variables.

6.3.2 Regression on the Residential Sector

We used the OLS model to do the regression, and the output of the regression is as follows:

OLS Regression Results

Dep. Variable:	Residential_Usage	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	23.79
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	1.21e-14
Time:	17:37:13	Log-Likelihood:	-400.58
No. Observations:	57	AIC:	821.2
Df Residuals:	47	BIC:	841.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2397.2771	541.796	4.425	0.000	1307.324	3487.230
Residential_Retail Price	36.4583	225.116	0.162	0.872	-416.416	489.332
CLDD	2595.1125	322.362	8.050	0.000	1946.604	3243.621
AWND	-643.4325	381.628	-1.686	0.098	-1411.169	124.304
HTDD	1943.3877	396.206	4.905	0.000	1146.325	2740.450
population	637.0272	367.731	1.732	0.090	-102.752	1376.807
solar-generation	831.3915	483.746	1.719	0.092	-141.780	1804.563
Summer	178.2910	192.076	0.928	0.358	-208.116	564.698
Fall	325.0916	169.937	1.913	0.062	-16.778	666.961
Winter	328.2478	172.261	1.906	0.063	-18.297	674.793

Omnibus:	3.103	Durbin-Watson:	2.319
Prob(Omnibus):	0.212	Jarque-Bera (JB):	2.196
Skew:	0.433	Prob(JB):	0.334
Kurtosis:	3.418	Cond. No.	31.2

R-squared and Adj. R-squared:

From in our model, the R-squared is 0.82, and the adjusted R-squared is 0.786, there is not much difference in them, so we are safe to say that our input variables are all valuable to the model, and they can explain 82% of the overall result.

Feature coefficients and Significance

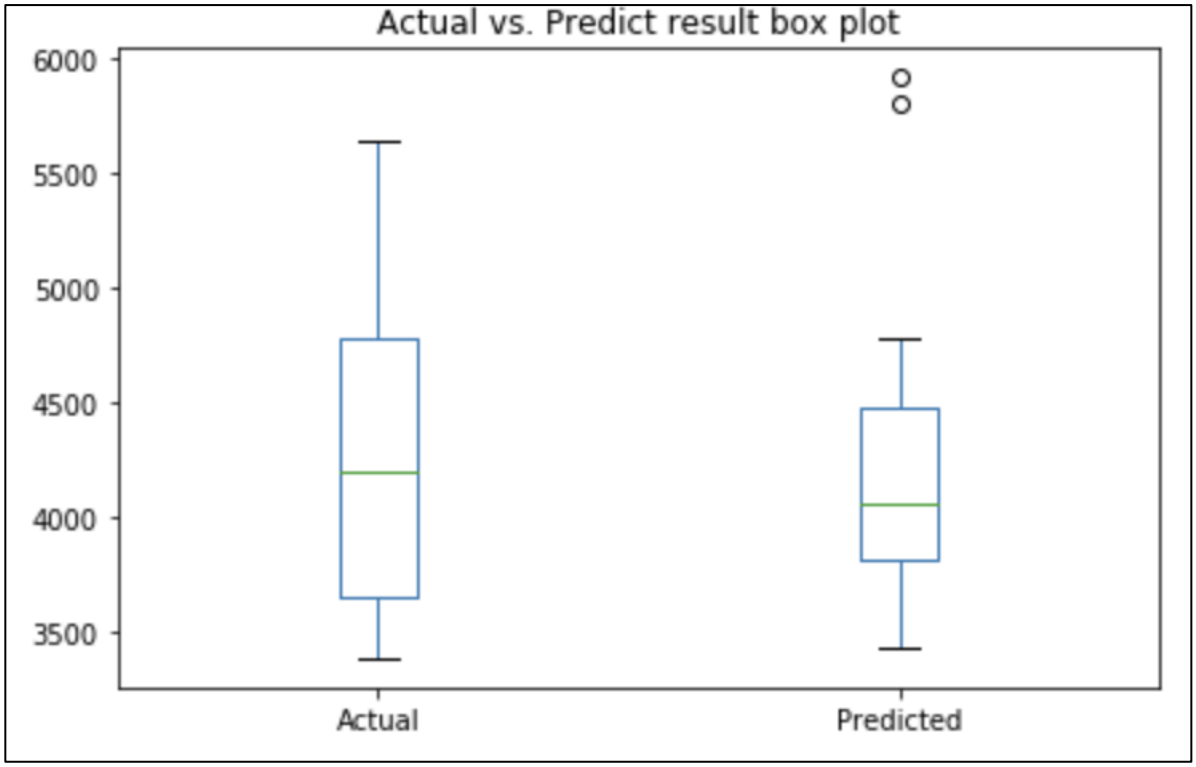
From the above table, we can say that CLDD and HTDD are significant features in the model. The logic is that with each increase in cooling degree days, there will be a corresponding 2595.1125 mkWh decrease in electricity consumption.

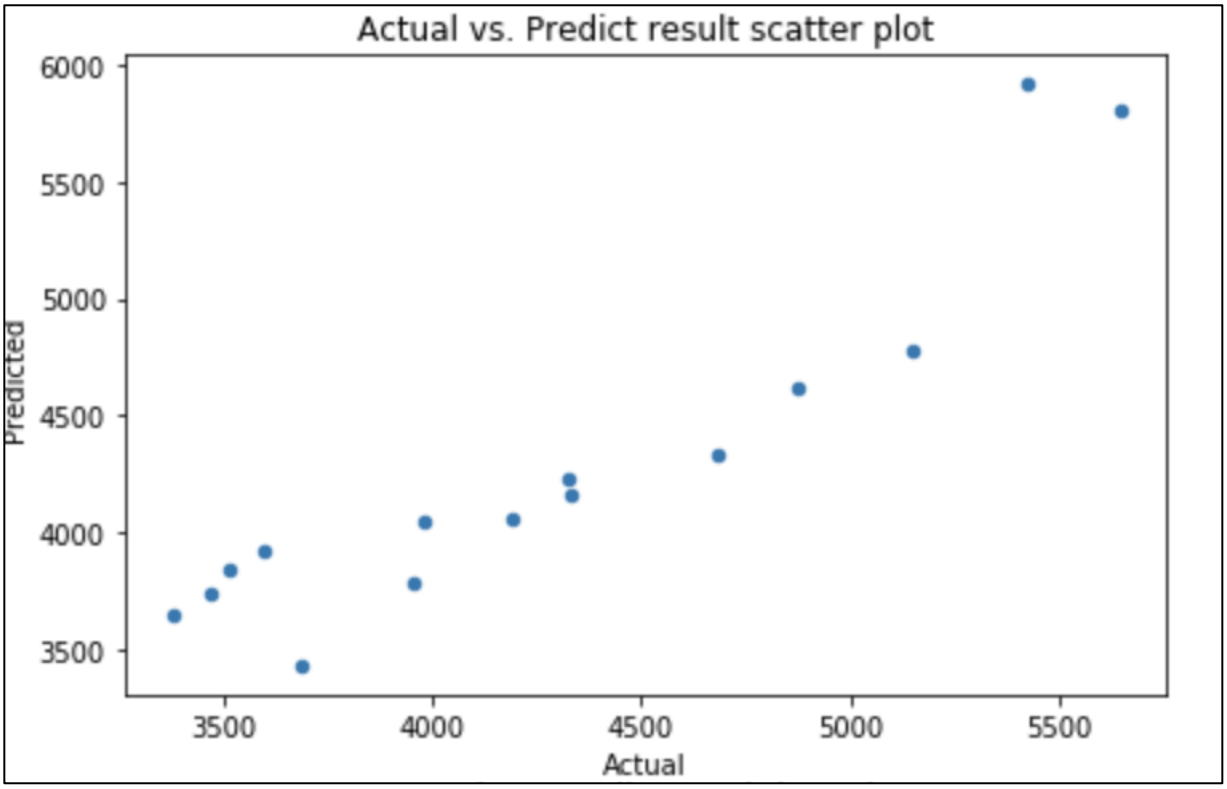
MAE, MSE, RMSE:

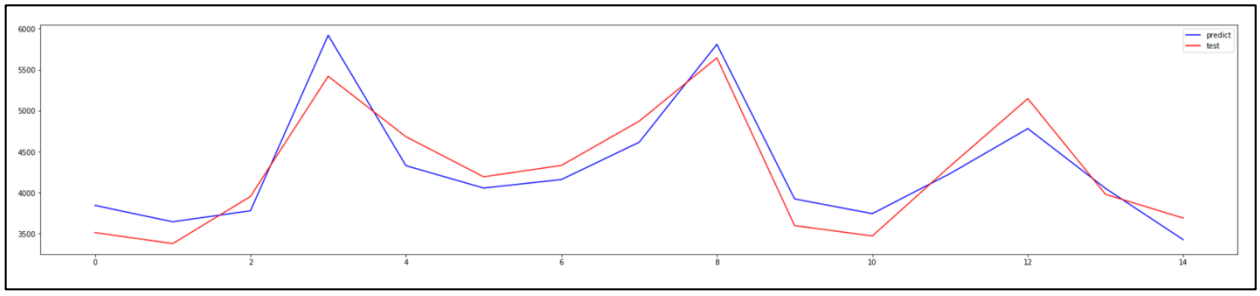
```
Performance Evaluation
Mean Absolute Error: 249.62738396461359
Mean Squared Error: 74872.42114601833
Root Mean Squared Error: 273.62825355949326
```

The MAE, MSE and RMSE metrics are used to evaluate how our trained model can help predict the future values. While a low error indicates that the model is representing the real scenario in a good manner, a high error does not necessarily mean that the model is useless. One possible reason for high error in the result may be due to some extreme values in the sample data that leads to extremely high error.

Actual vs Predicted:







In the third graph, the red line represents the value of real-world data, and the blue line represents the value that we predicted.

In all three graphs, we can tell that the model is making a good prediction of the test data. The trends in actual data and predicted value is quite close. The reason for the high error may be due to high electricity consumption in certain datapoints that makes the value of the error high. There are outliers in the dataset that are not representable by our linear model.

7. Florida

1. Commercial Consumption

OLS Regression Results

Dep. Variable:	Commercial_Usage	R-squared:	0.972
Model:	OLS	Adj. R-squared:	0.967
Method:	Least Squares	F-statistic:	183.9
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	1.57e-33
Time:	17:37:29	Log-Likelihood:	-359.24
No. Observations:	57	AIC:	738.5
Df Residuals:	47	BIC:	758.9
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6798.4006	144.374	47.089	0.000	6507.958	7088.843
Commercial_Retail Price	-407.1450	108.198	-3.763	0.000	-624.811	-189.479
CLDD	2491.3243	171.230	14.550	0.000	2146.854	2835.794
AWND	-155.7811	105.582	-1.475	0.147	-368.185	56.623
HTDD	895.0590	150.356	5.953	0.000	592.582	1197.536
population	-76.2007	220.006	-0.346	0.731	-518.797	366.395
solar-generation	224.8631	283.137	0.794	0.431	-344.736	794.462
Summer	11.8400	103.727	0.114	0.910	-196.833	220.513
Fall	134.1748	94.902	1.414	0.164	-56.742	325.092
Winter	-155.4811	96.927	-1.604	0.115	-350.473	39.511

Omnibus:	0.579	Durbin-Watson:	2.082
Prob(Omnibus):	0.749	Jarque-Bera (JB):	0.716
Skew:	-0.169	Prob(JB):	0.699
Kurtosis:	2.567	Cond. No.	31.4

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.972, adjusted R-squared is 0.967. This indicates that after adjusting for the number of predictors, 96.7% of the commercial energy consumption of Florida could be explained by the independent variables.

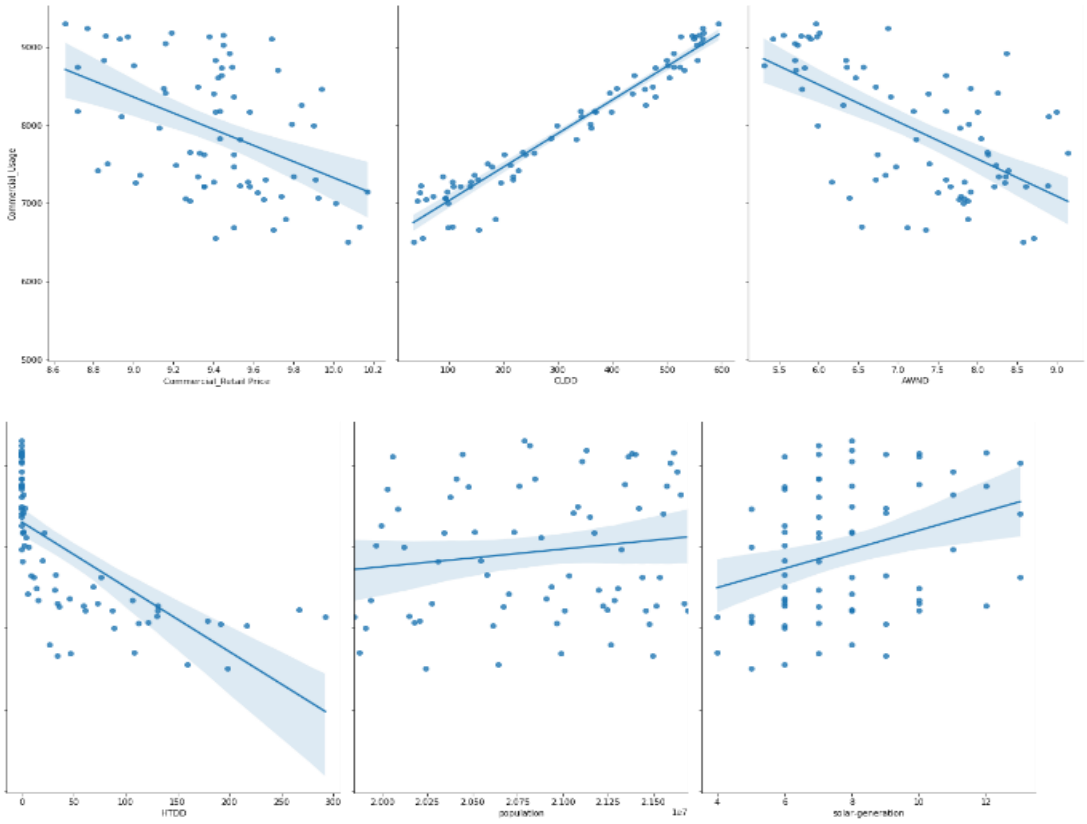
- Feature Coefficients and Significance
    - Among the independent variables, we could see that “Retail Price” has a significant negative impact on the commercial energy consumption: when retail price increase by 1 dollar, the monthly commercial energy consumption of Florida decreases by about 407 kilo-watthours.

    - In this regression, we don’t see very significant seasonal indicators, but CLDD and HTDD as two weather indicators are having very significant positive impact on commercial consumption: when indoors cooling days increase by 1, average monthly commercial consumption increase by 2491 kilo-watthours, when heating days increase by 1, average monthly commercial consumption increase by 895 kilo-watthours. This could because commercial consumption is taken up a lot by the indoor air conditioning in Florida.

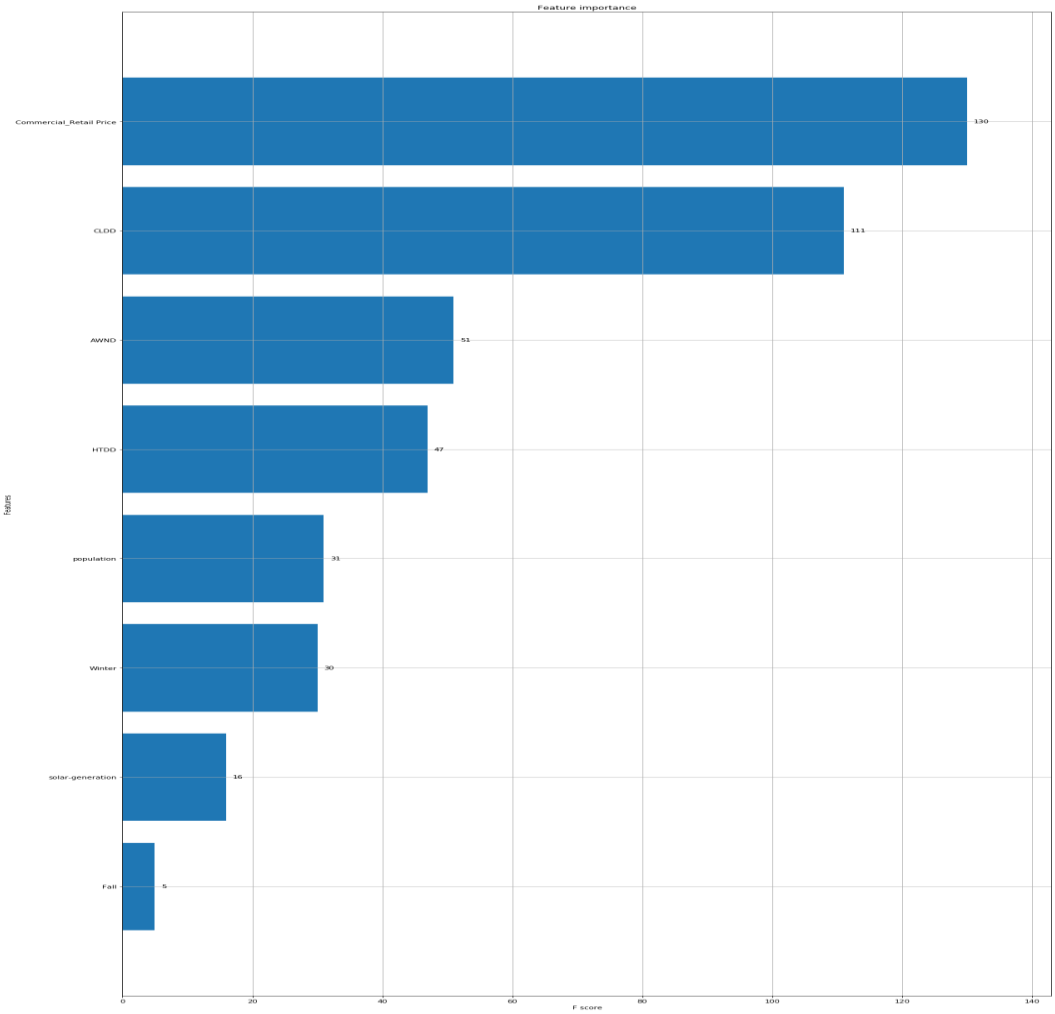
- MAE, MSE, RMSE
    - Mean Absolute Error: 108.19284151009245
    - Mean Squared Error: 17466.55178005011
    - Root Mean Squared Error: 132.1610826985392

  Though the model has a rather satisfactory prediction power, the MSE is very lar ge (17466).

- Feature Correlation

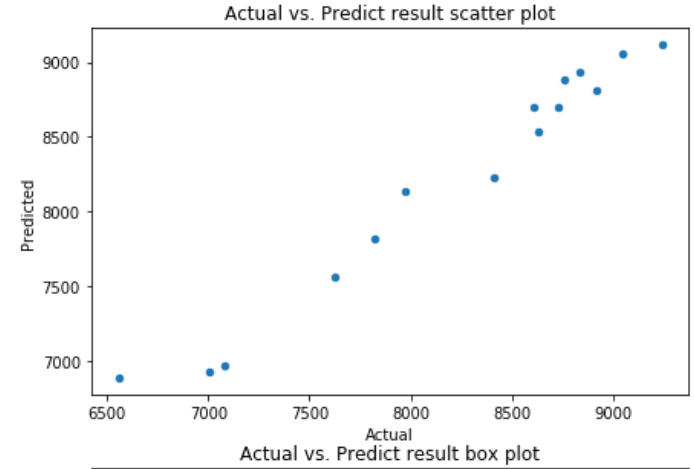


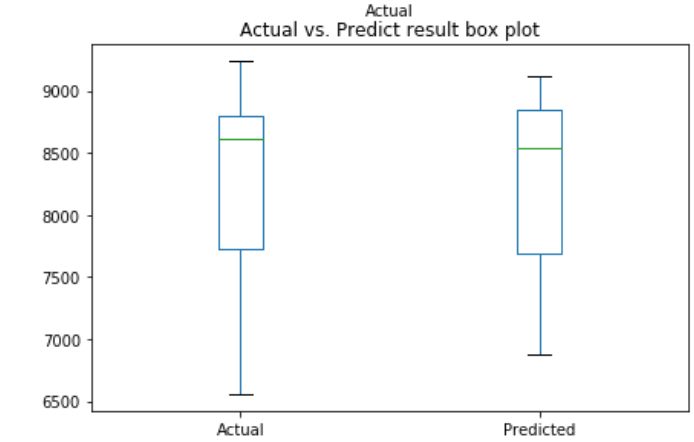
- Feature Importance

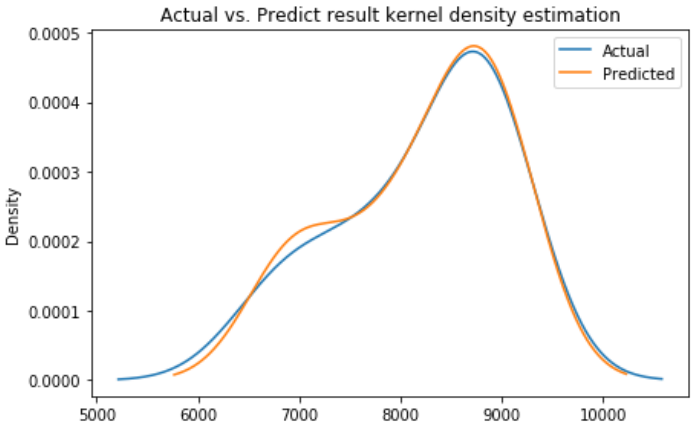


In all the features, top four features that contribute the most to the explanation of energy consumption are: Retail Price of electricity, CLDD, AWMD and HTDD, which are also the four significant variables we observe in the regression output.

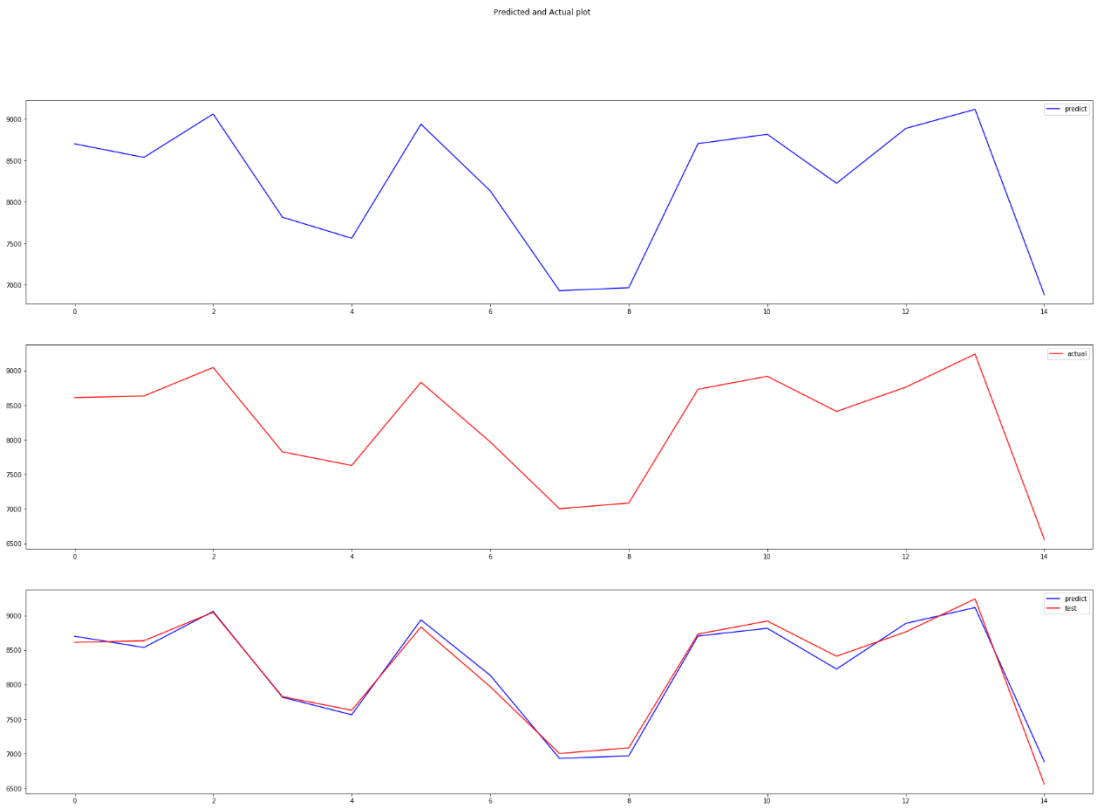
- Actual vs. Predict Visualization







- Actual vs. Prediction plot



From the visualization graphs we could see that the model is doing a relatively good job in predicting residential energy consumption.

2. Industrial Consumption

OLS Regression Results

Dep. Variable:	Industrial_Usage	R-squared:	0.843
Model:	OLS	Adj. R-squared:	0.813
Method:	Least Squares	F-statistic:	28.04
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	5.33e-16
Time:	17:37:39	Log-Likelihood:	-274.04
No. Observations:	57	AIC:	568.1
Df Residuals:	47	BIC:	588.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1338.1708	29.058	46.052	0.000	1279.714	1396.628
Industrial_Retail Price	-67.1081	27.207	-2.467	0.017	-121.841	-12.375
CLDD	238.9302	35.112	6.805	0.000	168.294	309.566
AWND	3.9842	23.721	0.168	0.867	-43.737	51.705
HTDD	120.8163	33.424	3.615	0.001	53.575	188.057
population	44.3119	41.181	1.076	0.287	-38.533	127.157
solar-generation	-128.0090	54.973	-2.329	0.024	-238.601	-17.417
Summer	-8.7638	23.156	-0.378	0.707	-55.348	37.820
Fall	-54.4330	19.381	-2.809	0.007	-93.423	-15.443
Winter	-80.9744	20.398	-3.970	0.000	-122.010	-39.939

Omnibus:	0.211	Durbin-Watson:	2.351
Prob(Omnibus):	0.900	Jarque-Bera (JB):	0.053
Skew:	0.074	Prob(JB):	0.974
Kurtosis:	2.990	Cond. No.	25.4

- R-squared, Adjusted R-squared
  The value of R-squared is 0.843, adjusted R-squared is 0.813. This indicates that after adjusting for the number of predictors, 81.3% of the industrial energy consumption of Florida could be explained by the independent variables.

- Feature Coefficients and Significance
  - Among the independent variables, we could see that “Retail Price” has a rather significant negative impact on the industrial energy consumption: when retail price increase by 1 dollar, the monthly industrial energy consumption of Florida decreases by about 67 kilo-watthours. The coefficient is significant at 5% level of confidence level.

  - CLDD and HTDD as two weather indicators are having very significant positive impact on industrial consumption: when indoors cooling days increase by 1, average monthly industrial consumption increase by about 239 kilo-watthours, when heating days increase by 1, average monthly industrial consumption increase by about 121 kilo-watthours.

    Noticeably, cooling days has a greater impact on industrial energy consumption compared with heating days, this could because of Florida’s

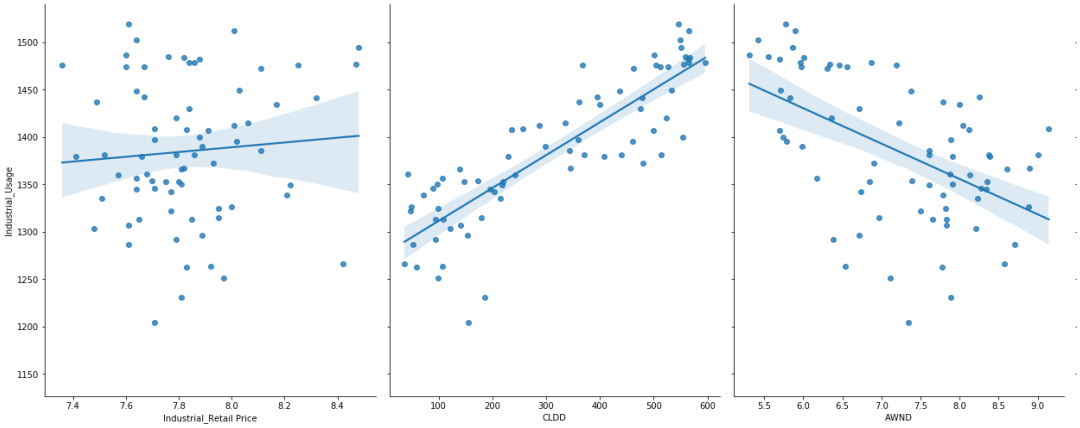
geographic and climate characteristics, that average temperature is pretty high and heating is not as needed in winter.

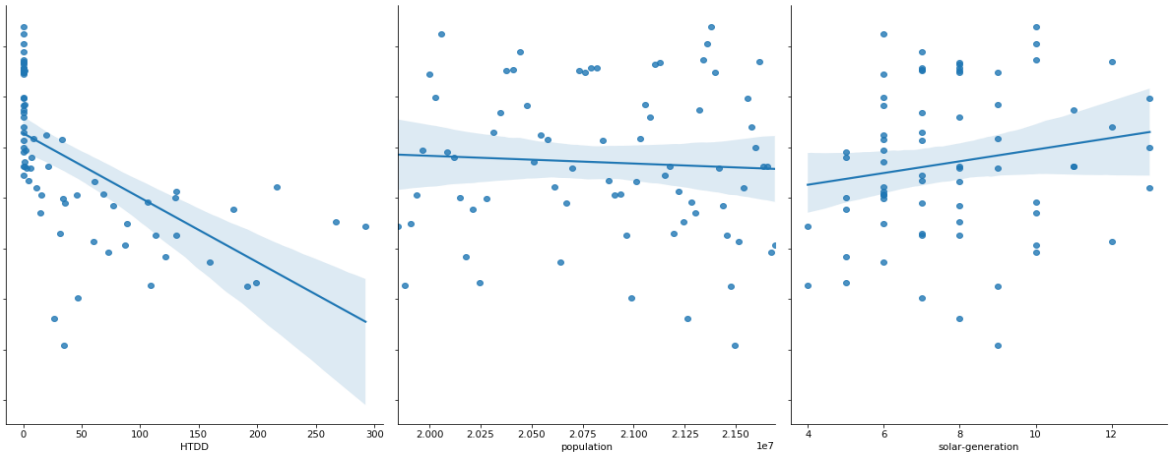
- Fall and Winter as two seasonal dummy variables are being significant at 1% level of confidence. They indicate that, compared with spring, the industrial energy consumption in fall is about 54 kilo-watthours lower per month, and that in winter, the industrial energy consumption is about 81 kilo-watthours lower per month.

  The reason behind this phenomenon is not hard to figure: Florida is using air conditioning the most in its two hottest seasons, spring and summer. While in fall and winter, temperature drops compared to spring and therefore electricity consumption drops with less use of air conditioning.

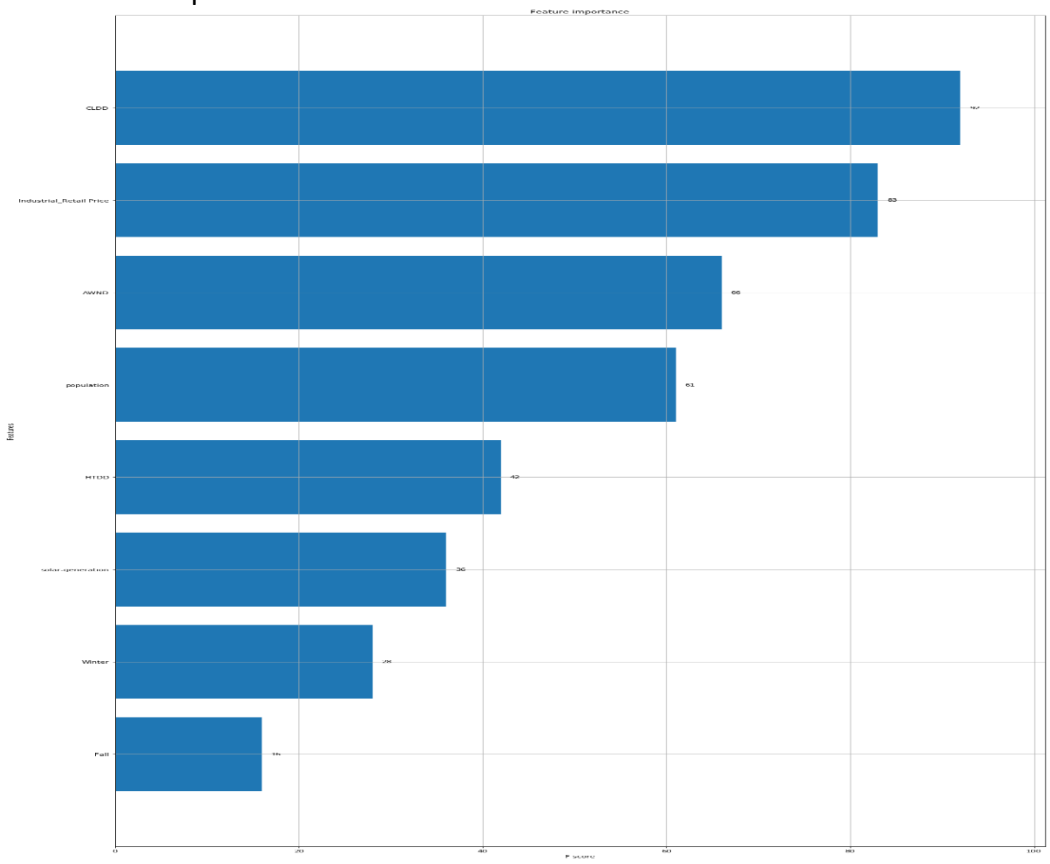
- MAE, MSE, RMSE
  - Mean Absolute Error: 26.904630167428042
  - Mean Squared Error: 896.9368603301275
  - Root Mean Squared Error: 29.948904159086148

- Correlation Analysis



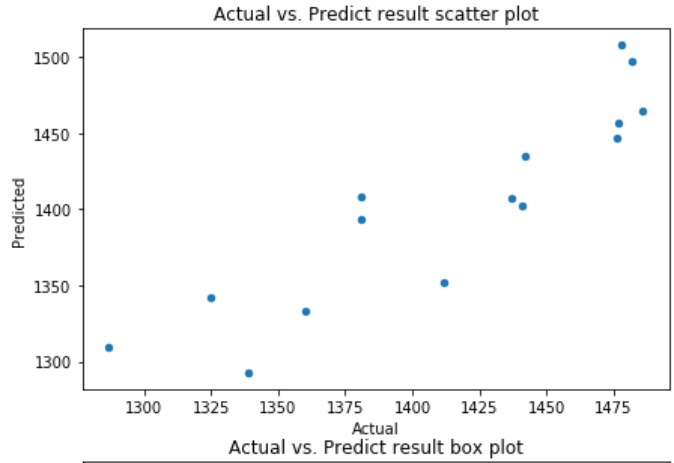


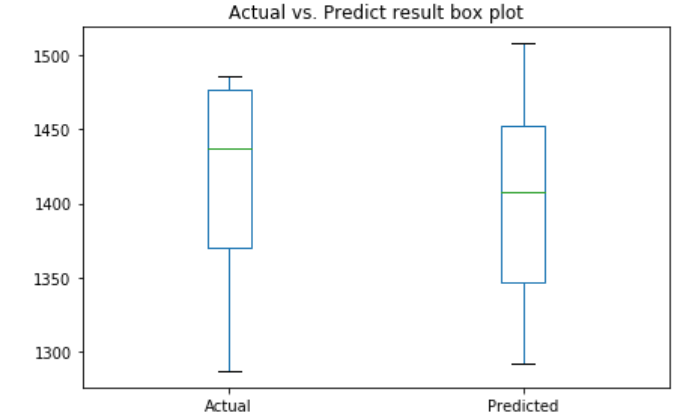
- Feature Importance

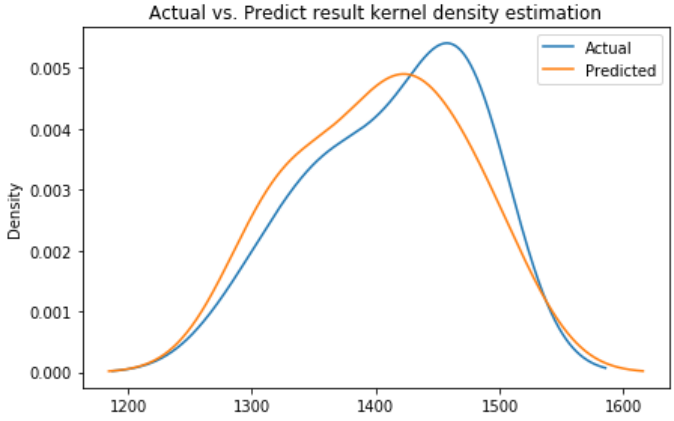


In all the features, top three features that contribute the most to the explanation of energy consumption are: CLDD, AWND and Retail Price of electricity. Cooling days impact the energy consumption the most greatly, while retail price of electricity is not as important.

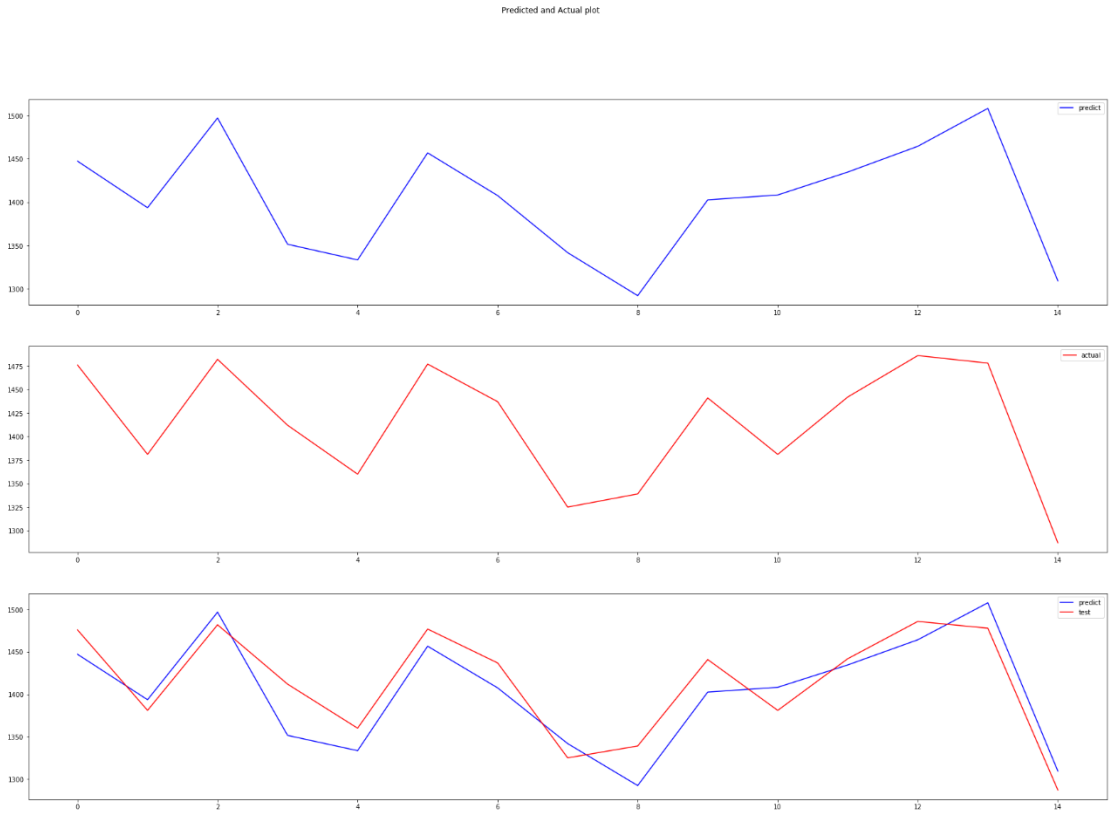
- Actual vs. Predict Visualization







- Actual vs. Prediction plot



- From the visualization graphs we could see that the model is doing a relatively good job in predicting residential energy consumption.

3. Residential Consumption

OLS Regression Results

Dep. Variable:	Residential_Usage	R-squared:	0.968
Model:	OLS	Adj. R-squared:	0.962
Method:	Least Squares	F-statistic:	156.9
Date:	Mon, 20 Jul 2020	Prob (F-statistic):	5.77e-32
Time:	17:37:48	Log-Likelihood:	-409.08
No. Observations:	57	AIC:	838.2
Df Residuals:	47	BIC:	858.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5977.2946	317.923	18.801	0.000	5337.717	6616.872
Residential_Retail Price	-365.9701	261.345	-1.400	0.168	-891.728	159.788
CLDD	6736.0438	390.622	17.244	0.000	5950.214	7521.873
AWND	-327.6492	254.748	-1.286	0.205	-840.136	184.838
HTDD	3659.9294	357.204	10.246	0.000	2941.328	4378.531
population	641.3696	482.185	1.330	0.190	-328.662	1611.401
solar-generation	112.3008	676.527	0.166	0.869	-1248.696	1473.298
Summer	360.7180	249.130	1.448	0.154	-140.468	861.904
Fall	428.4540	228.905	1.872	0.067	-32.044	888.952
Winter	371.4773	227.934	1.630	0.110	-87.066	830.021

Omnibus:	2.860	Durbin-Watson:	2.159
Prob(Omnibus):	0.239	Jarque-Bera (JB):	1.993
Skew:	0.417	Prob(JB):	0.369
Kurtosis:	3.379	Cond. No.	30.4

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.968, adjusted R-squared is 0.962. This indicates that after adjusting for the number of predictors, 96.2% of the residential energy consumption of Florida could be explained by the independent variables.

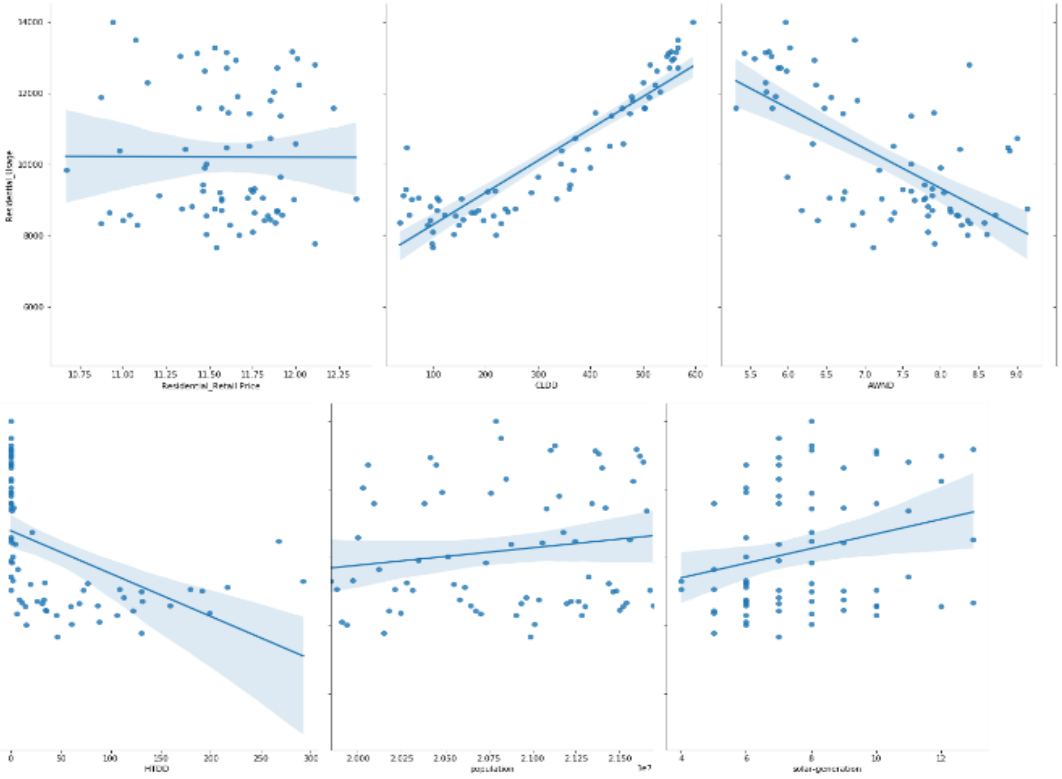
- Feature Coefficients and Significance
    - Among the independent variables, we could see that “Retail Price” has a negative impact on the residential energy consumption: when retail price increase by 1 dollar, the monthly residential energy consumption of Florida decreases by about 366 kilo-watthours. However, this coefficient is not significant for residential consumption in Florida, which suggests that residential use of electricity is not that sensitive to retail prices as in other states.

    - In this regression, we don’t see very significant seasonal indicators, but CLDD and HTDD as two weather indicators are having very significant positive impact on residential consumption: when indoors cooling days increase by 1, average monthly residential consumption increase by 6736 kilo-watthours, when heating days increase by 1, average monthly residential consumption increase by 3659 kilo-watthours.

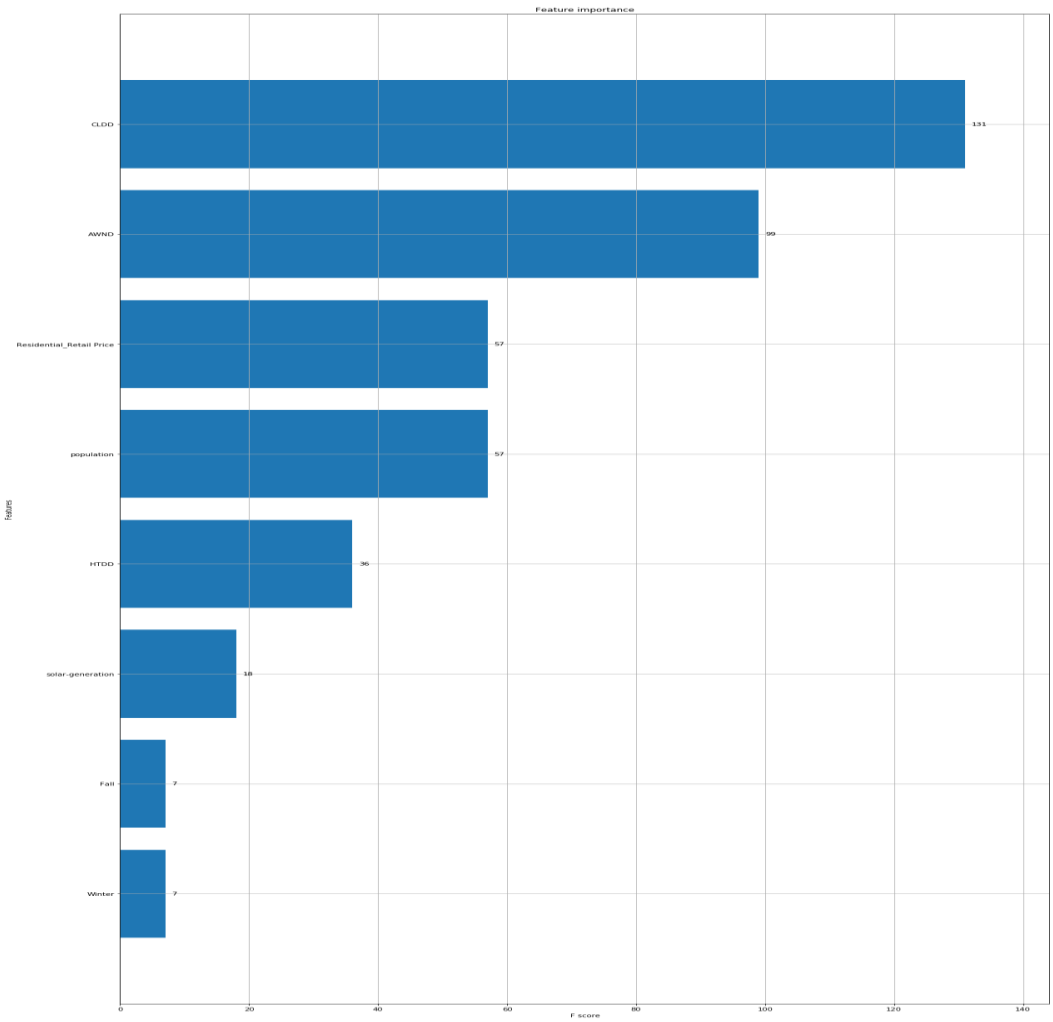
Noticeably, cooling days has a greater impact on residential energy consumption compared with heating days, this could because of Florida's geographic and climate characteristics, that average temperature is pretty high and heating is not as needed in winter.

- MAE, MSE, RMSE
  - Mean Absolute Error: 362.9004323059893
  - Mean Squared Error: 184585.34019612707
  - Root Mean Squared Error: 429.63396071089056

- Correlation Analysis

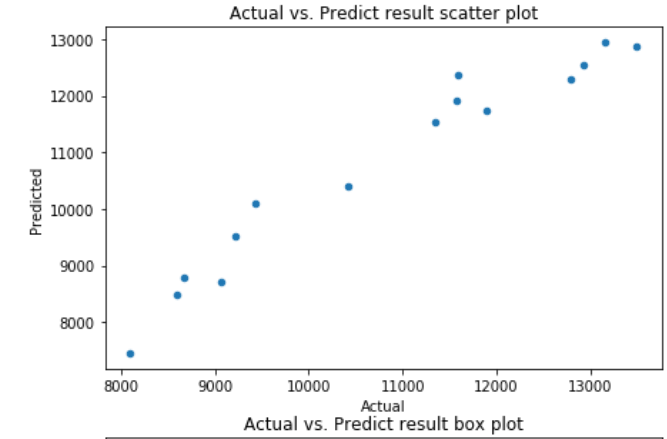


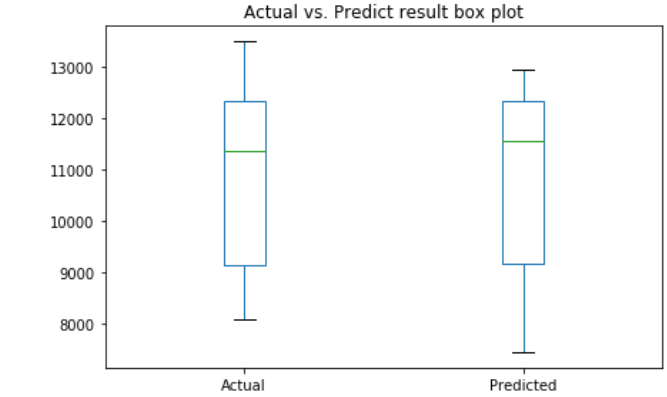
- Feature Importance

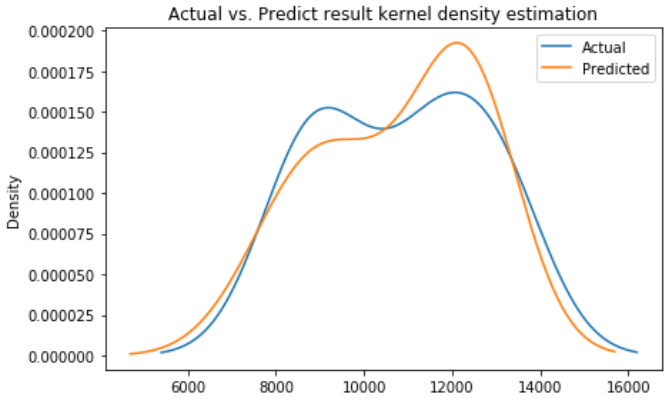


In all the features, top three features that contribute the most to the explanation of energy consumption are: CLDD, AWND and Retail Price of electricity. Cooling days impact the energy consumption the most greatly, while retail price of electricity is not as important.

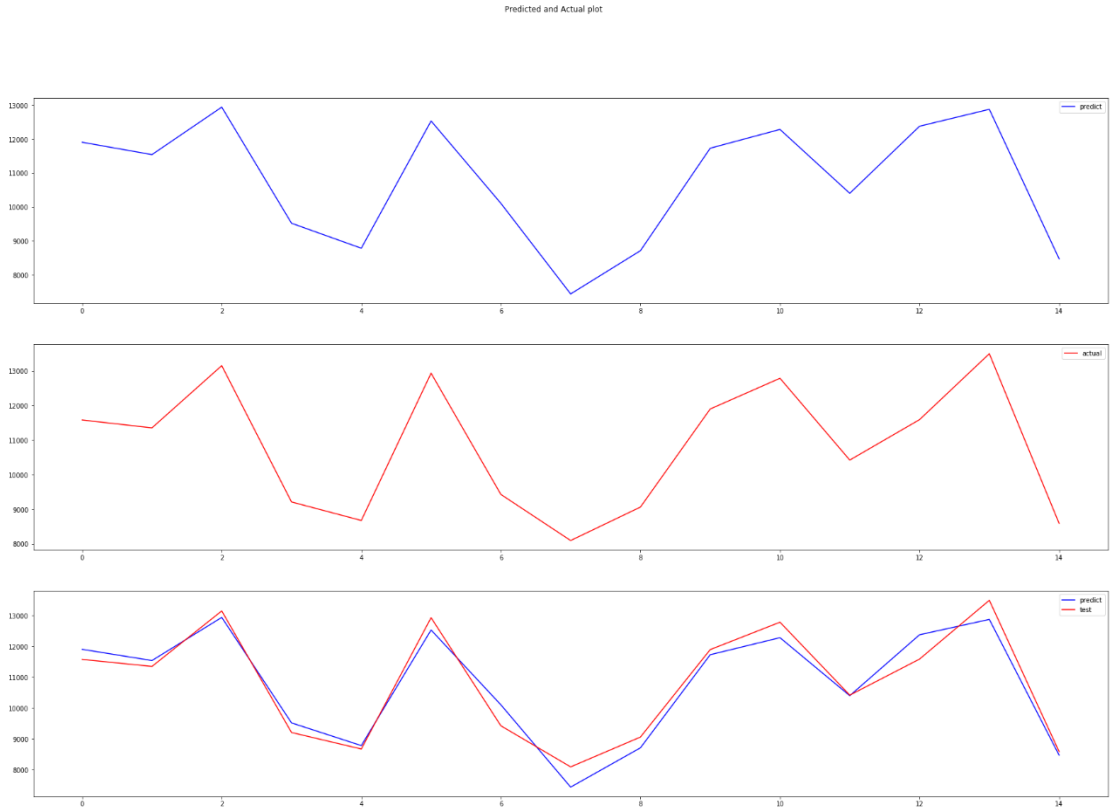
- Actual vs. Predict Visualization







- Actual vs. Prediction plot



From the visualization graphs we could see that the model is doing a relatively good job in predicting residential energy consumption.

8. Wyoming

1. VIF

   We calculated the VIF of all numeric independent variables in order to check for multicollinearity. The calculated VIF is as follows:

```
Commercial_Retail Price          3.391521
CLDD                             3.507672
TAVG                           238.418383
AWND                             1.327667
HTDD                           221.186600
area                        173137.417073
population                       2.609052
solar-generation                      NaN
dtype: float64
```

```
Residential_Retail Price          5.768860
CLDD                              3.192110
TAVG                            244.496380
AWND                              1.399142
HTDD                            216.950097
area                         136342.231978
population                        2.105796
solar-generation                       NaN
dtype: float64
```

```
Industrial_Retail Price           1.362012
CLDD                              3.014997
TAVG                            244.207343
AWND                              1.511519
HTDD                            225.626930
area                          66822.986450
population                        1.049866
solar-generation                       NaN
dtype: float64
```

We could observe that variable TAVG, HTDD and area have a very large VIF, indicating that there might be multicollinearity among the variables. Also, solar-generation is having nan values. We try to remove TAVG, area and solar-generation to eliminate multicollinearity, and the VIF after removal is as follows:

```
const                        167028.308496
Commercial_Retail Price           3.382126
CLDD                              2.855039
AWND                              1.258292
HTDD                              4.015370
population                        2.607332
dtype: float64
```

```
const                         61200.130906
Industrial_Retail Price           1.326043
CLDD                              2.338100
AWND                              1.477080
HTDD                              2.544285
population                        1.049816
dtype: float64
```

```
const                        136172.815961
Residential_Retail Price          5.609868
CLDD                              2.372345
AWND                              1.355017
HTDD                              6.106802
population                        2.066438
dtype: float64
```

2. Commercial Consumption

OLS Regression Results

Dep. Variable:	Commercial_Usage	R-squared:	0.726
Model:	OLS	Adj. R-squared:	0.681
Method:	Least Squares	F-statistic:	15.94
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	3.60e-11
Time:	16:30:01	Log-Likelihood:	-211.44
No. Observations:	57	AIC:	440.9
Df Residuals:	48	BIC:	459.3
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	345.4770	13.202	26.168	0.000	318.932	372.022
Commercial_Retail Price	-63.9255	12.152	-5.260	0.000	-88.360	-39.491
CLDD	-5.4037	13.737	-0.393	0.696	-33.024	22.217
AWND	2.4306	5.986	0.406	0.687	-9.606	14.467
HTDD	7.4434	11.752	0.633	0.529	-16.186	31.072
population	-0.4794	7.297	-0.066	0.948	-15.151	14.192
Summer	17.4579	7.810	2.235	0.030	1.754	33.162
Fall	-3.5352	4.281	-0.826	0.413	-12.143	5.073
Winter	3.2387	5.551	0.583	0.562	-7.923	14.401

Omnibus:	6.228	Durbin-Watson:	2.128
Prob(Omnibus):	0.044	Jarque-Bera (JB):	5.535
Skew:	-0.555	Prob(JB):	0.0628
Kurtosis:	4.047	Cond. No.	20.9

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.726, adjusted R-squared is 0.681. This indicates that after adjusting for the number of predictors, 68.1% of the commercial energy consumption of Wyoming could be explained by the independent variables.

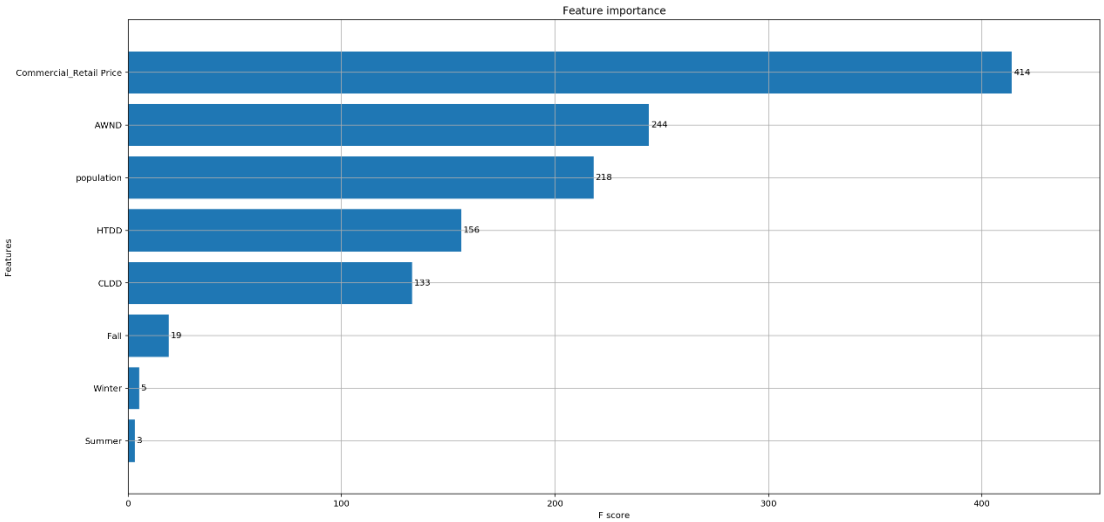
- Feature Coefficients and Significance
    - Among the independent variables, we could see that “Retail Price” has a significant negative impact on the commercial energy consumption: when retail price increase by 1 dollar, the monthly commercial energy consumption of Wyoming decreases by about 63.9 kilo-watthours. The price sensitivity in residential sector is higher than that in residential or commercial sectors.

    - Among the independent variables, we don’t see very significant climate indicators, but we could see that “Summer” as a seasonal dummy variable is significant at 5% level of confidence, which indicates that, compared with spring, the commercial energy consumption in summer is about 17.45 kilo-watthours higher per month.

- MAE, MSE, RMSE
    - Mean Absolute Error: 14.261350053845982
    - Mean Squared Error: 278.5177938264

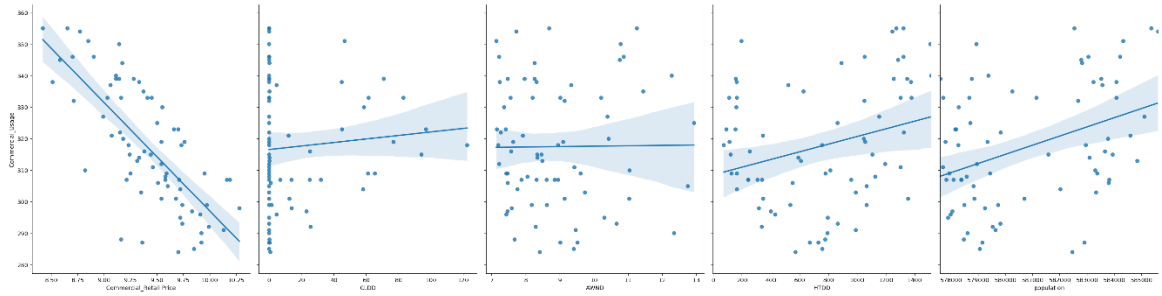
- Root Mean Squared Error: 16.68885238194646

- Feature Importance

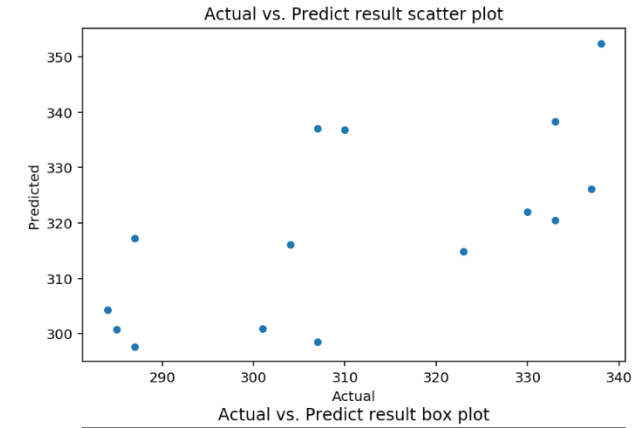


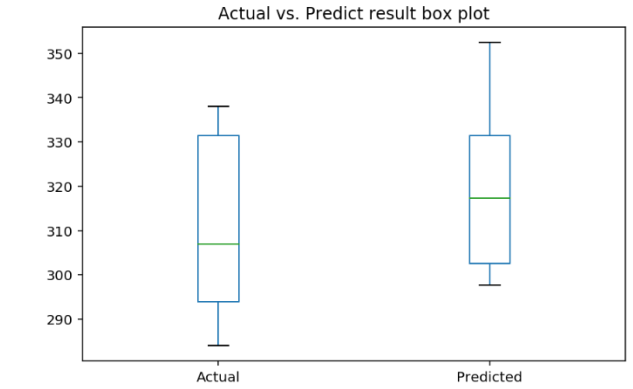
In all the features, top four features that contribute the most to the explanation of energy consumption are: Retail Price of electricity, AWND, Population and HTDD. Here, CLDD does not matter that much in commercial consumption in Wyoming. This could because Wyoming is in the northern part of the country and therefore more sensitive to heating days but not cooling days.

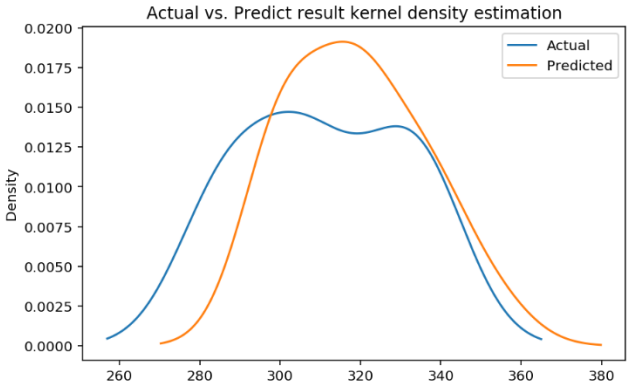
- Correlation Analysis



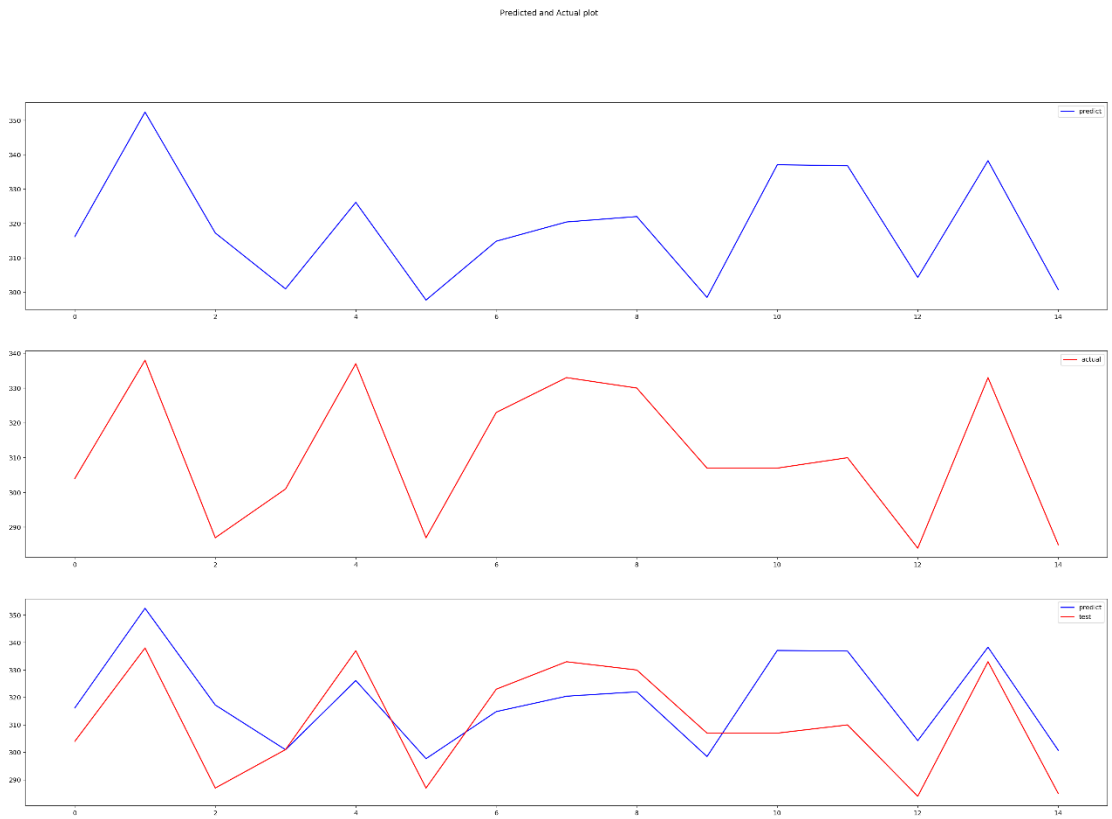
- Actual vs. Predict Visualization







- Actual vs. Prediction plot



From the visualization graphs we could see that the model is not doing a very good job in predicting commercial energy consumption, which resonates with our R squared results.

3. Industrial Consumption

OLS Regression Results

Dep. Variable:	Industrial_Usage	R-squared:	0.261
Model:	OLS	Adj. R-squared:	0.138
Method:	Least Squares	F-statistic:	2.122
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	0.0517
Time:	16:30:57	Log-Likelihood:	-264.63
No. Observations:	57	AIC:	547.3
Df Residuals:	48	BIC:	565.6
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	869.9480	26.374	32.985	0.000	816.920	922.976
Industrial_Retail Price	-37.0837	21.920	-1.692	0.097	-81.157	6.989
CLDD	41.5799	33.147	1.254	0.216	-25.066	108.225
AWND	-27.3064	17.256	-1.582	0.120	-62.001	7.389
HTDD	15.3236	30.487	0.503	0.618	-45.974	76.621
population	2.1742	11.876	0.183	0.856	-21.704	26.052
Summer	-12.9337	20.742	-0.624	0.536	-54.638	28.770
Fall	6.1776	11.324	0.546	0.588	-16.590	28.945
Winter	4.6867	13.885	0.338	0.737	-23.231	32.605

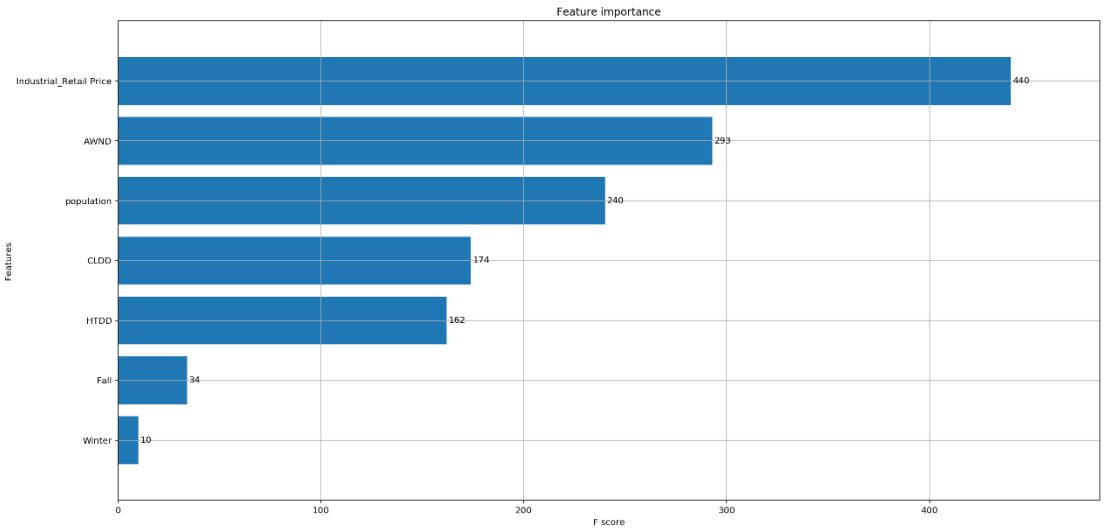
Omnibus:	0.051	Durbin-Watson:	2.136
Prob(Omnibus):	0.975	Jarque-Bera (JB):	0.235
Skew:	-0.020	Prob(JB):	0.889
Kurtosis:	2.688	Cond. No.	17.3

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.261, adjusted R-squared is 0.138. This indicates that after adjusting for the number of predictors, 92.6% of the industrial energy consumption of Wyoming could be explained by the independent variables.

- Feature Coefficients and Significance
    - Among the independent variables, we could see that “Retail Price” has a negative impact on the industrial energy consumption, which is significant at 10% level of confidence. This indicates that when retail price increase by 1 dollar, the monthly industrial energy consumption of Wyoming decreases by about 37.1 kilo-watthours.

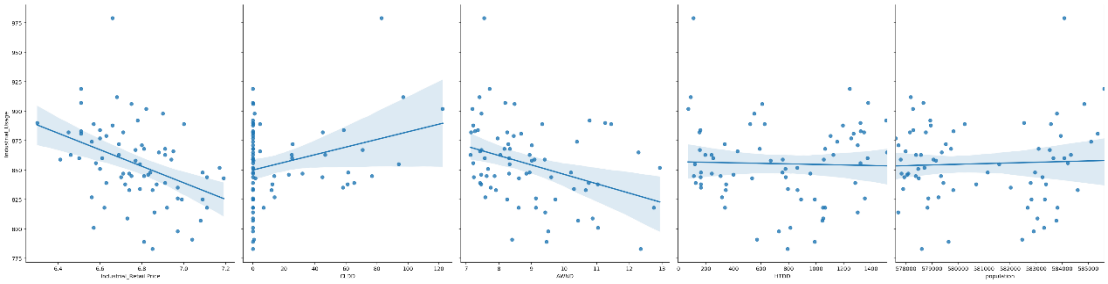
- MAE, MSE, RMSE
    - Mean Absolute Error: 29.157571127622123
    - Mean Squared Error: 1520.2074910686983
    - Root Mean Squared Error: 38.9898383052392

- Feature Importance

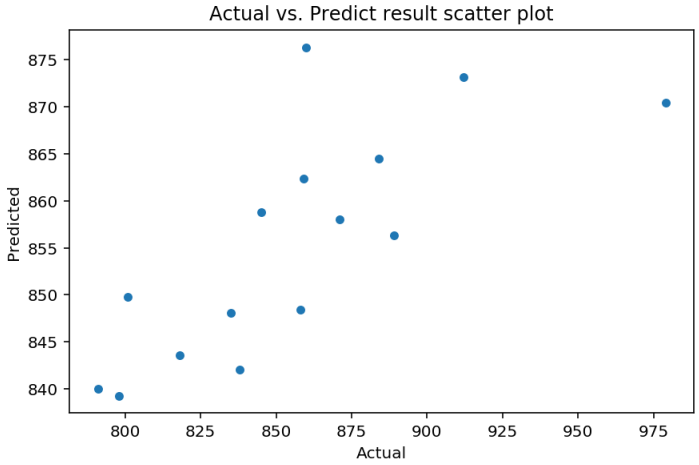


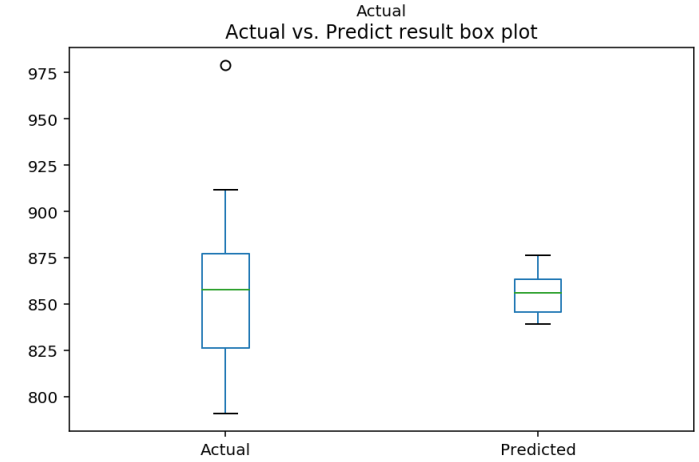
In all the features, top three features that contribute the most to the explanation of energy consumption are: Retail Price of electricity, AWND and Population. Here, CLDD does not matter that much in industrial consumption in Wyoming.

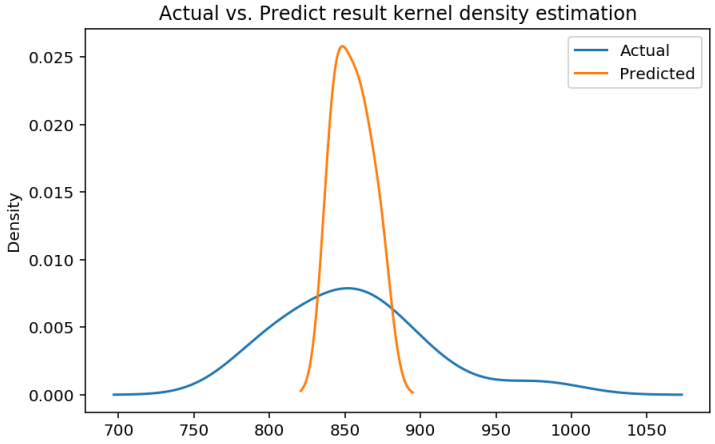
- Correlation Analysis



- Actual vs. Predict Visualization







- Actual vs. Prediction plot

- From the visualization graphs we could see that the model is doing a relatively poor job in predicting industrial energy consumption.

4. Residential Consumption

OLS Regression Results

Dep. Variable:	Residential_Usage	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.922
Method:	Least Squares	F-statistic:	83.85
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	1.48e-25
Time:	16:31:54	Log-Likelihood:	-225.97
No. Observations:	57	AIC:	469.9
Df Residuals:	48	BIC:	488.3
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	182.9433	22.055	8.295	0.000	138.598	227.288
Residential_Retail Price	-41.5535	20.307	-2.046	0.046	-82.384	-0.723
CLDD	77.4668	16.853	4.597	0.000	43.581	111.353
AWND	-7.1573	7.868	-0.910	0.368	-22.976	8.661
HTDD	139.6302	19.098	7.311	0.000	101.232	178.028
population	-10.2846	8.618	-1.193	0.239	-27.612	7.043
Summer	10.3353	10.002	1.033	0.307	-9.776	30.447
Fall	-14.8487	5.790	-2.565	0.014	-26.490	-3.208
Winter	20.9374	7.007	2.988	0.004	6.849	35.026

Omnibus:	4.871	Durbin-Watson:	2.326
Prob(Omnibus):	0.088	Jarque-Bera (JB):	3.986
Skew:	-0.624	Prob(JB):	0.136
Kurtosis:	3.344	Cond. No.	26.7

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.933, adjusted R-squared is 0.922. This indicates that after adjusting for the number of predictors, 92.2% of the residential energy consumption of Wyoming could be explained by the independent variables.

- Feature Coefficients and Significance
    - Among the independent variables, we could see that “Retail Price” has a negative impact on the residential energy consumption, which is significant at 5% level of confidence. This indicates that when retail price increase by 1 dollar, the monthly industrial energy consumption of Wyoming decreases by about 41.6 kilo-watthours.

    - CLDD and HTDD as two weather indicators are having very significant positive impact on residential consumption: when indoors cooling days increase by 1, average monthly commercial consumption increase by 77 kilo-watthours, when heating days increase by 1, average monthly commercial consumption increase by 139 kilo-watthours.

      Noticeably, heating days has a greater impact on industrial energy consumption compared with cooling days, this could because of

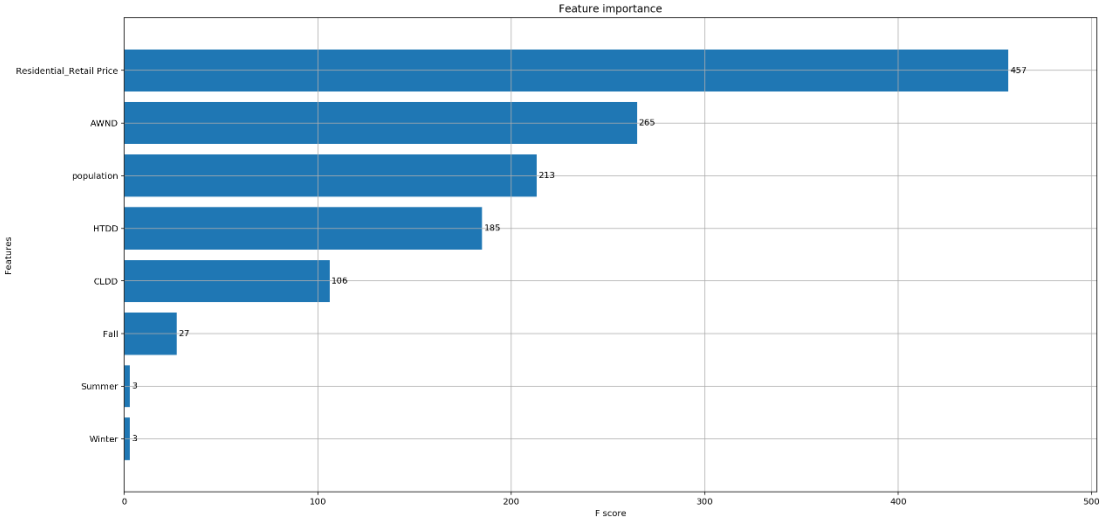
Wyoming's geographic and climate characteristics, that average temperature is pretty low and cooling is not as needed in summer as heating is needed in winter.

- Fall and Winter as two seasonal dummy variables are being significant at 5% and 1% level of confidence respectively. They indicate that, compared with spring, the residential energy consumption in fall is about 14.8 kilo-watthours lower per month, and that in winter, the industrial energy consumption is about 20.9 kilo-watthours higher per month.

  The reason behind this phenomenon is not hard to figure: Wyoming is consuming the most energy in its two coldest seasons, spring and winter. Therefore, the energy consumption in winter raises with the frequent use of heating devices.

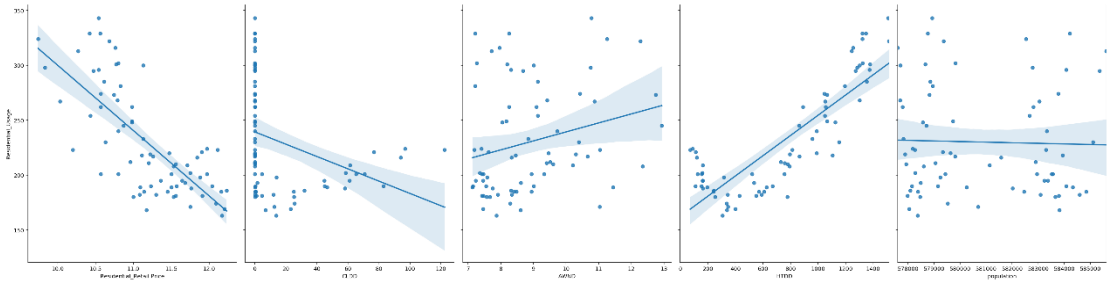
- MAE, MSE, RMSE
  - Mean Absolute Error: 10.2209363045067
  - Mean Squared Error: 147.74575513047782
  - Root Mean Squared Error: 12.155071169288883

- Feature Importance

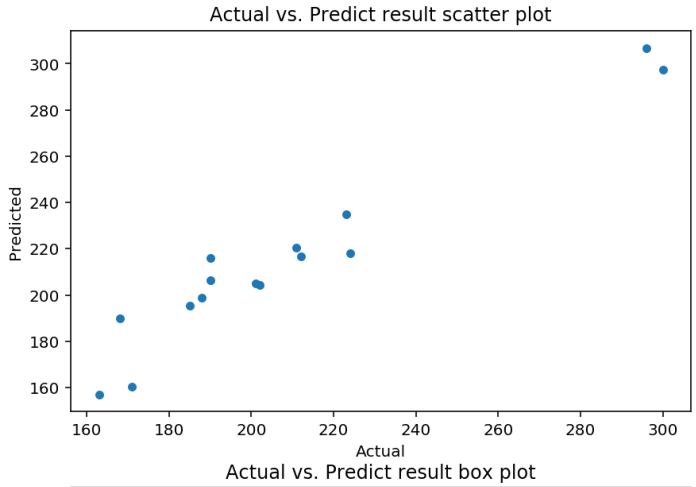


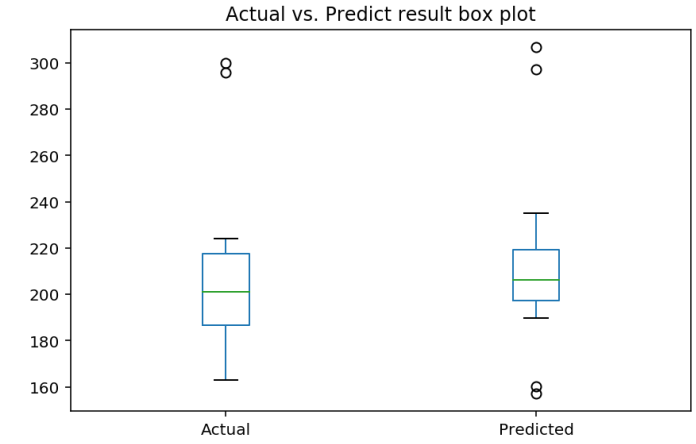
In all the features, top four features that contribute the most to the explanation of energy consumption are: Retail Price of electricity, AWND, Population and HTDD. Here, CLDD does not matter that much in commercial consumption in Wyoming. This could because Wyoming is in the northern part of the country and therefore more sensitive to heating days but not cooling days.

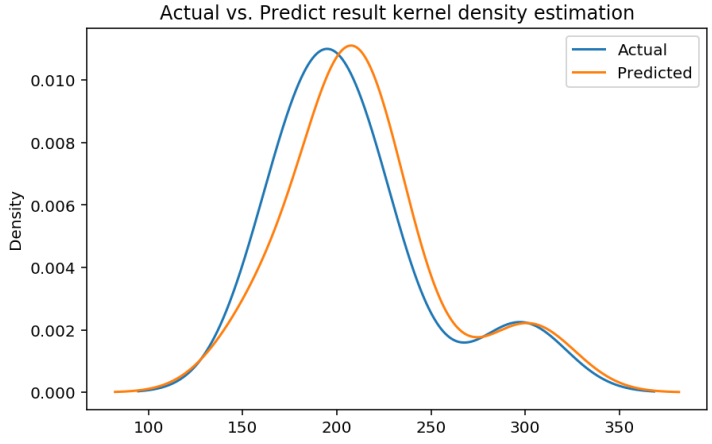
- Correlation Analysis



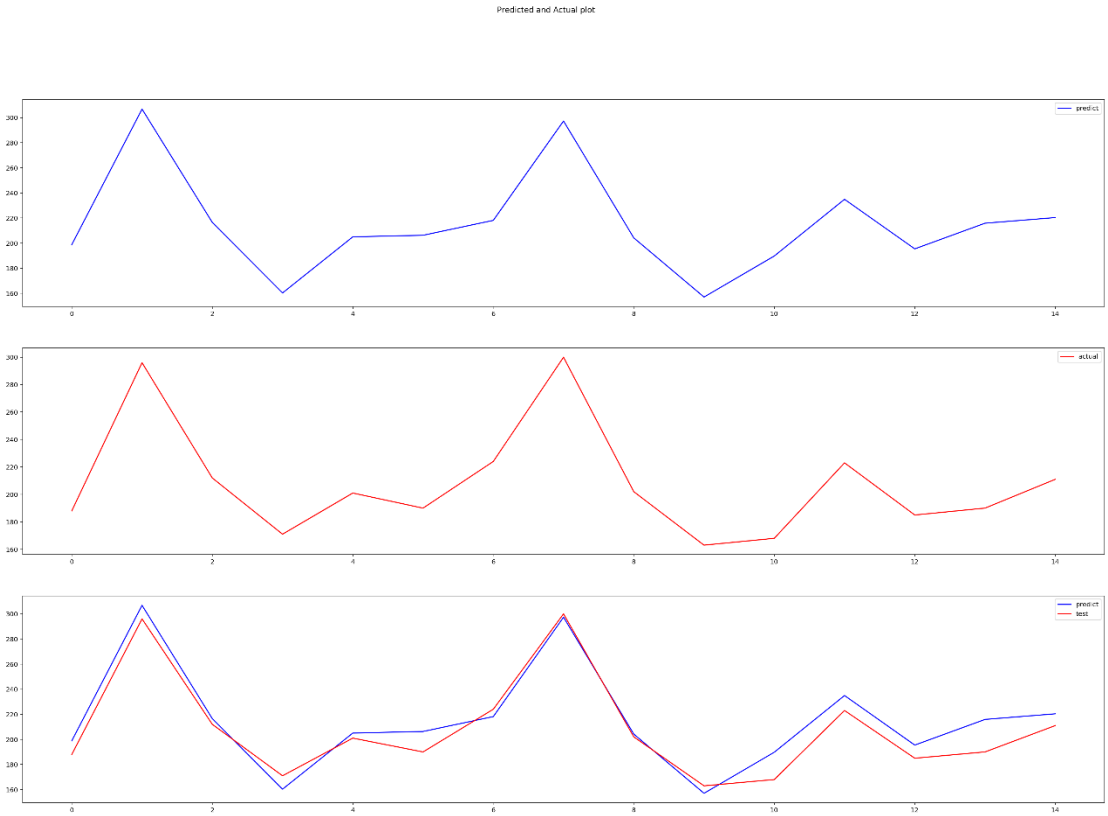
- Actual vs. Predict Visualization







- Actual vs. Prediction plot



From the visualization graphs we could see that prediction of the model pretty poor, which is consistent with the R squared results in regression outputs.

9. California

1. VIF
   We calculated the VIF of all numeric independent variables in order to check for multicollinearity. The calculated VIF is as follows:

   Commercial Sector

```
Commercial_Retail Price        5.981140
CLDD                          89.067753
TAVG                         606.242420
AWND                           1.775001
HTDD                         277.715541
area                       92447.756911
population                     3.314668
solar-generation               4.409602
dtype: float64
```

   Residential Sector

```
Residential_Retail Price        1.938744
CLDD                           88.582742
TAVG                          607.227841
AWND                            1.380519
HTDD                          272.559490
area                        93165.697311
population                      3.430611
solar-generation                4.029965
dtype: float64
```

Industrial Sector

```
Industrial_Retail Price        5.364034
CLDD                          86.656522
TAVG                         602.299125
AWND                           1.915826
HTDD                         275.202040
area                       92388.505006
population                     3.333238
solar-generation               4.475700
dtype: float64
```

We could observe that variable TAVG, HTDD and area have a very large VIF, indicating that there might be multicollinearity among the variables. We try to remove TAVG and area to eliminate multicollinearity, and the VIF after removal is as follows:

Commercial Sector

```
const                       60382.632268
Commercial_Retail Price         5.920372
CLDD                            5.698946
AWND                            1.696462
HTDD                            5.374792
population                      3.296876
solar-generation                4.306940
dtype: float64
```

Residential Sector

```
const                        62585.391426
Residential_Retail Price         1.915932
CLDD                             5.249797
AWND                             1.324660
HTDD                             5.724195
population                       3.418188
solar-generation                 3.946806
dtype: float64
```

Industrial Sector

```
const                       60583.192682
Industrial_Retail Price         5.344297
CLDD                            5.083920
AWND                            1.848062
HTDD                            5.487456
population                      3.313536
solar-generation                4.388743
dtype: float64
```

All VIF values are smaller than 10, indicating that there is no multicollinearity among the independent variables now.

2. **Commercial Consumption**

We regress the data with OLS model, and the output of regression is as follows:

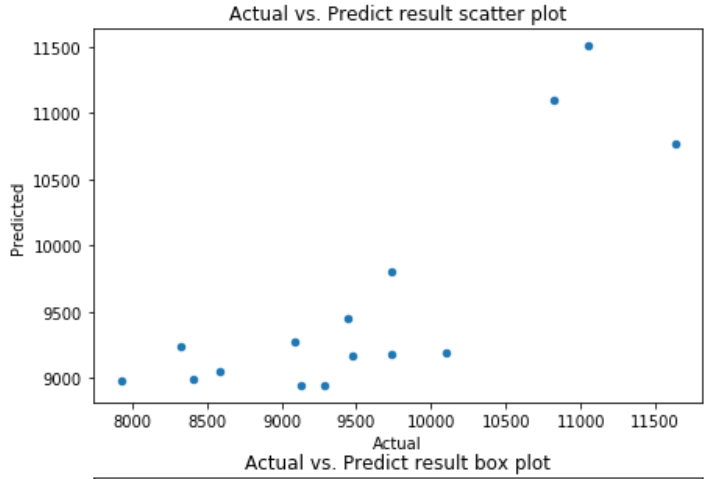
```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Commercial_Usage   R-squared:                       0.631
Model:                           OLS   Adj. R-squared:                  0.561
Method:                Least Squares   F-statistic:                     8.946
Date:               Tue, 21 Jul 2020   Prob (F-statistic):           1.05e-07
Time:                       13:49:59   Log-Likelihood:                -440.99
No. Observations:                 57   AIC:                             902.0
Df Residuals:                     47   BIC:                             922.4
Df Model:                          9
Covariance Type:           nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                     9196.1840    630.291     14.590      0.000    7928.202    1.05e+04
Commercial_Retail Price   1296.2333    943.579      1.374      0.176    -602.003    3194.470
CLDD                      2716.3914    937.184      2.898      0.006     831.020    4601.763
AWND                      -478.1224    556.108     -0.860      0.394   -1596.867     640.623
HTDD                       697.1246    864.154      0.807      0.424   -1041.328    2435.577
population                -149.4744    568.455     -0.263      0.794   -1293.059     994.110
solar-generation          -707.5136    740.955     -0.955      0.345   -2198.124     783.097
Summer                    -645.4726    461.615     -1.398      0.169   -1574.122     283.177
Fall                      -352.0545    482.096     -0.730      0.469   -1321.906     617.797
Winter                    -437.1099    426.552     -1.025      0.311   -1295.223     421.003
==============================================================================
Omnibus:                        8.284   Durbin-Watson:                   1.666
Prob(Omnibus):                  0.016   Jarque-Bera (JB):               12.534
Skew:                          -0.395   Prob(JB):                      0.00190
Kurtosis:                       5.157   Cond. No.                         27.1
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

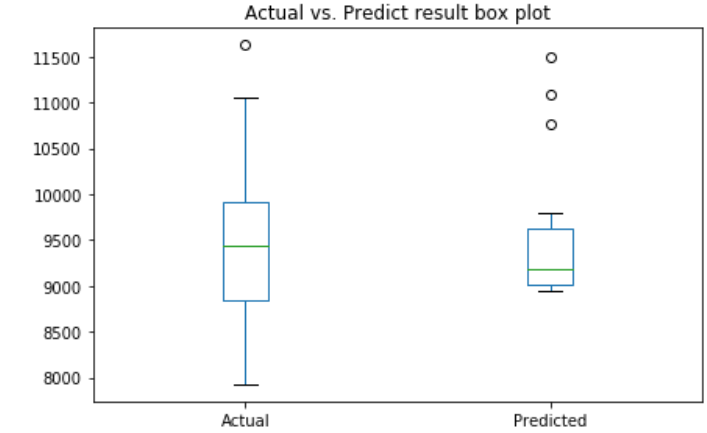
- R-squared, Adjusted R-squared
    - The value of R-squared is 0.631, adjusted R-squared is 0.561. This indicates that after adjusting for the number of predictors, 56.1% of the commercial energy consumption of California could be explained by the independent variables.
    - CLDD as a weather indicator is having very significant positive impact on residential consumption: when indoors cooling days increase by 1, average monthly commercial consumption increase by 2716.4 kilo-watthours.

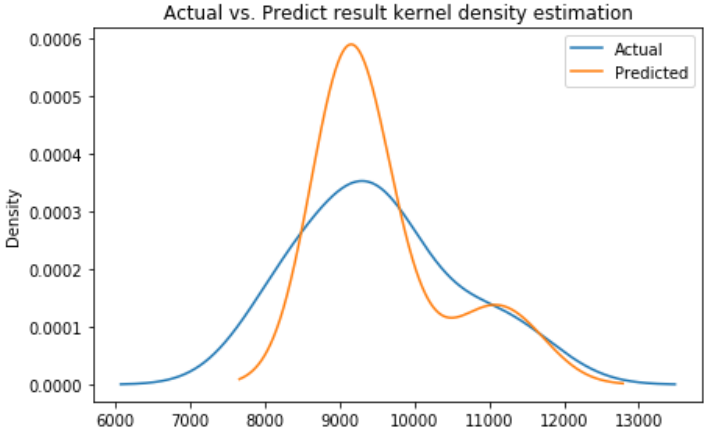
- MAE, MSE, RMSE

```
Performance Evaluation
Mean Absolute Error: 476.50229160348573
Mean Squared Error: 328613.89826780366
Root Mean Squared Error: 573.2485484218897
```

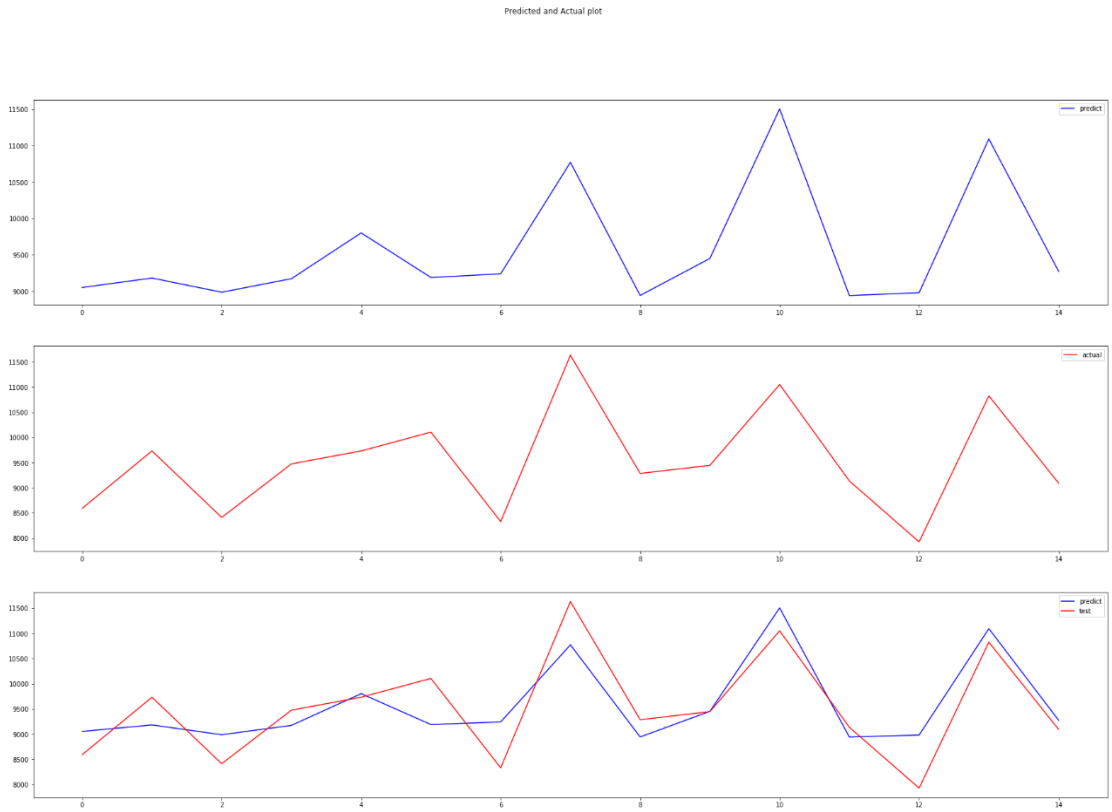
- Actual vs. Prediction Visualization


Actual vs. Predict result scatter plot
Predicted
Actual
11500
11000
10500
10000
9500
9000
8000
8500
9000
9500
10000
10500
11000
11500


Actual vs. Predict result box plot
11500
11000
10500
10000
9500
9000
8500
8000
Actual
Predicted


Actual vs. Predict result kernel density estimation
Density
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
6000
7000
8000
9000
10000
11000
12000
13000
Actual
Predicted

- Actual vs. Prediction plot



3. Industrial Consumption

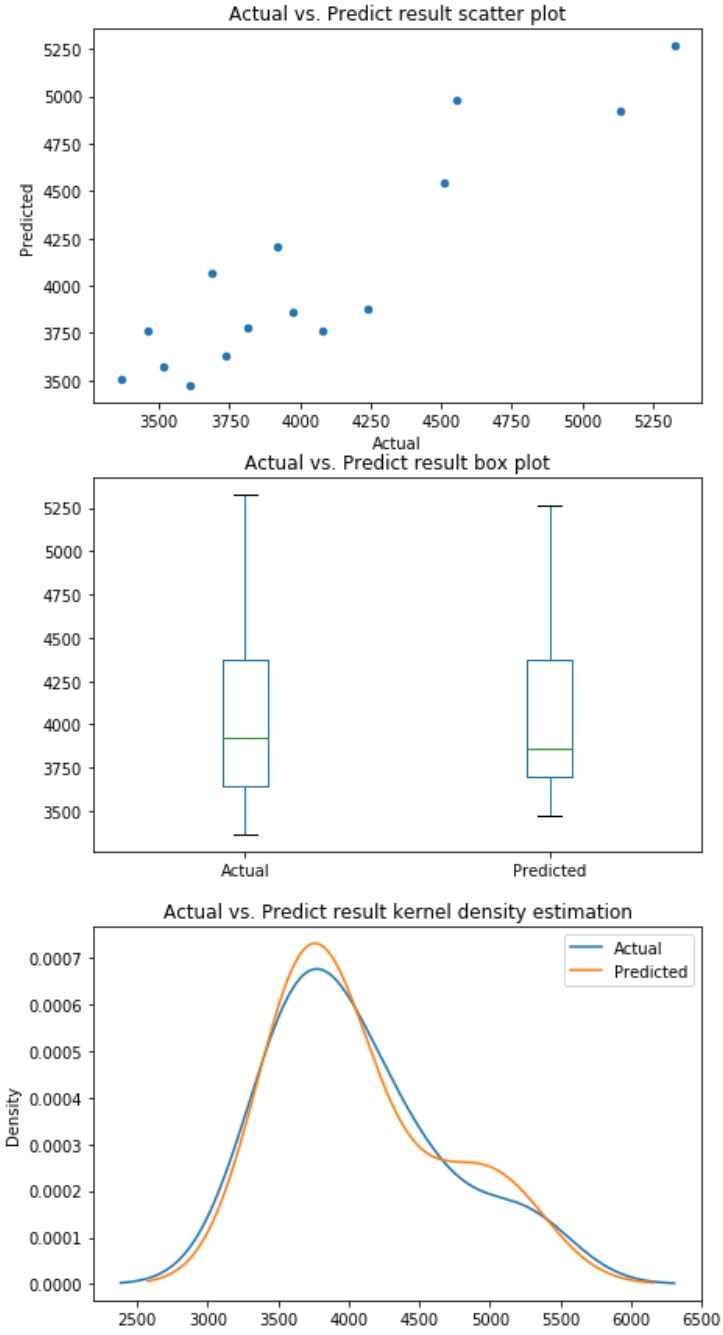
```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Industrial_Usage   R-squared:                       0.860
Model:                           OLS   Adj. R-squared:                  0.833
Method:                Least Squares   F-statistic:                     32.02
Date:               Tue, 21 Jul 2020   Prob (F-statistic):           4.00e-17
Time:                       13:50:02   Log-Likelihood:                -382.71
No. Observations:                 57   AIC:                             785.4
Df Residuals:                     47   BIC:                             805.8
Df Model:                          9
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                     4178.7432    225.825     18.504      0.000    3724.443    4633.044
Industrial_Retail Price    463.3500    281.893      1.644      0.107    -103.745    1030.445
CLDD                      1100.1805    303.791      3.622      0.001     489.033    1711.329
AWND                       178.2128    200.504      0.889      0.379    -225.149     581.575
HTDD                      -378.7294    311.157     -1.217      0.230   -1004.696     247.237
population                -192.2997    203.028     -0.947      0.348    -600.739     216.139
solar-generation          -611.5277    265.999     -2.299      0.026   -1146.649     -76.407
Summer                    -224.1763    190.956     -1.174      0.246    -608.330     159.977
Fall                        -0.3408    183.938     -0.002      0.999    -370.377     369.695
Winter                    -138.3312    150.186     -0.921      0.362    -440.466     163.804
==============================================================================
Omnibus:                        0.115   Durbin-Watson:                   2.214
Prob(Omnibus):                  0.944   Jarque-Bera (JB):                0.127
Skew:                           0.091   Prob(JB):                        0.938
Kurtosis:                       2.857   Cond. No.                         25.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.860, adjusted R-squared is 0.833. This indicates that after adjusting for the number of predictors, 83.3% of the industrial energy consumption of California could be explained by the independent variables. This performs relative better than commercial sectors.
    - CLDD as a weather indicator is having very significant positive impact on industrial consumption: when indoors cooling days increase by 1, average monthly industrial consumption increase by 1100.2 kilo-watthours,

- MAE, MSE, RMSE

```
Performance Evaluation
Mean Absolute Error: 197.66721769206285
Mean Squared Error: 56214.810934014524
Root Mean Squared Error: 237.09662784192972
```

- Actual vs. Predict Visualization

- 

- Actual vs. Prediction plot

4. Residential Consumption

OLS Regression Results

Dep. Variable:	Residential_Usage	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.742
Method:	Least Squares	F-statistic:	18.89
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	7.99e-13
Time:	13:50:06	Log-Likelihood:	-449.59
No. Observations:	57	AIC:	919.2
Df Residuals:	47	BIC:	939.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5628.0163	755.411	7.450	0.000	4108.326	7147.707
Residential_Retail Price	-877.2436	687.994	-1.275	0.209	-2261.310	506.822
CLDD	7812.0001	987.769	7.909	0.000	5824.865	9799.135
AWND	-1232.4699	649.927	-1.896	0.064	-2539.954	75.014
HTDD	2593.4821	1007.736	2.574	0.013	566.179	4620.785
population	-70.5321	678.580	-0.104	0.918	-1435.660	1294.596
solar-generation	510.7335	725.443	0.704	0.485	-948.670	1970.137
Summer	-918.8411	510.124	-1.801	0.078	-1945.077	107.395
Fall	0.6678	421.090	0.002	0.999	-846.457	847.793
Winter	656.8153	471.174	1.394	0.170	-291.064	1604.695

Omnibus:	22.433	Durbin-Watson:	2.043
Prob(Omnibus):	0.000	Jarque-Bera (JB):	66.055
Skew:	0.980	Prob(JB):	4.53e-15
Kurtosis:	7.896	Cond. No.	26.1

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.783, adjusted R-squared is 0.742. This indicates that after adjusting for the number of predictors, 74.2% of the industrial energy consumption of California could be explained by the independent variables.

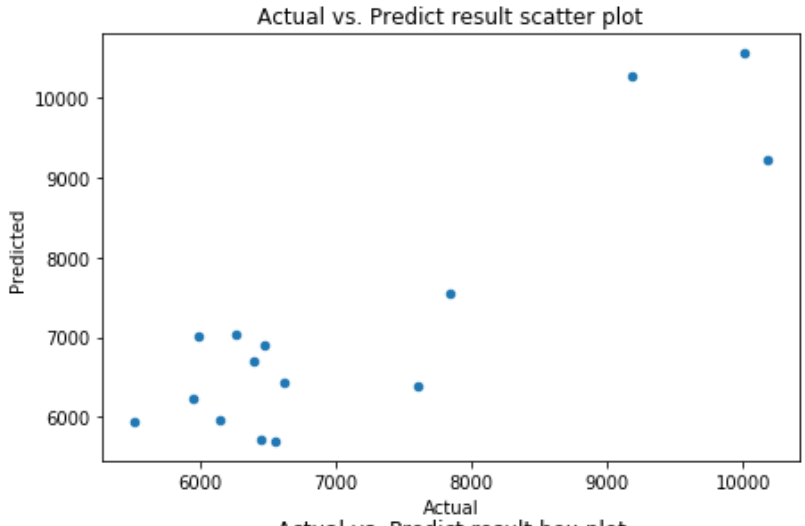
    - CLDD and HTDD as two weather indicators are having very significant positive impact on residential consumption: when indoors cooling days increase by 1, average monthly residential consumption increase by 7812.0 kilo-watthours, when heating days increase by 1, average monthly residential consumption increase by 2593.4 kilo-watthours.

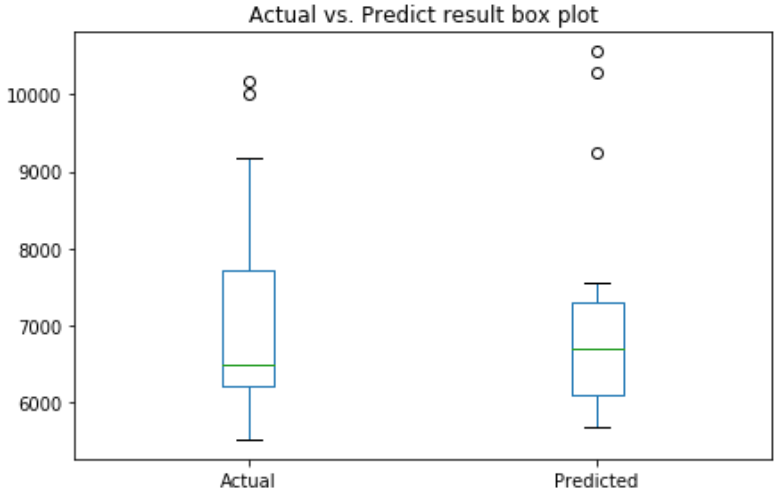
      Noticeably, cooling days has a greater impact on industrial energy consumption compared with heating days, this could because of California's geographic and climate characteristics, that average temperature is pretty high and heating is not as needed in winter as cooling is needed in summer.

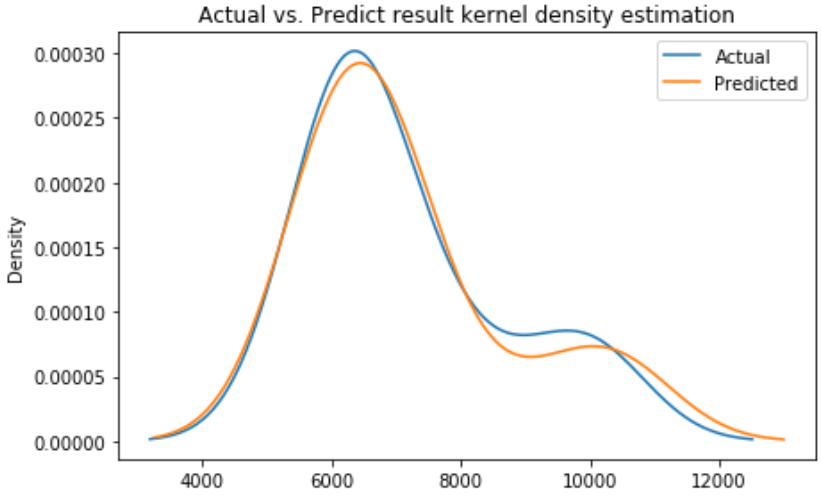
- MAE, MSE, RMSE

```
Performance Evaluation
Mean Absolute Error: 616.7428590053264
Mean Squared Error: 495157.3207022466
Root Mean Squared Error: 703.674158046355
```

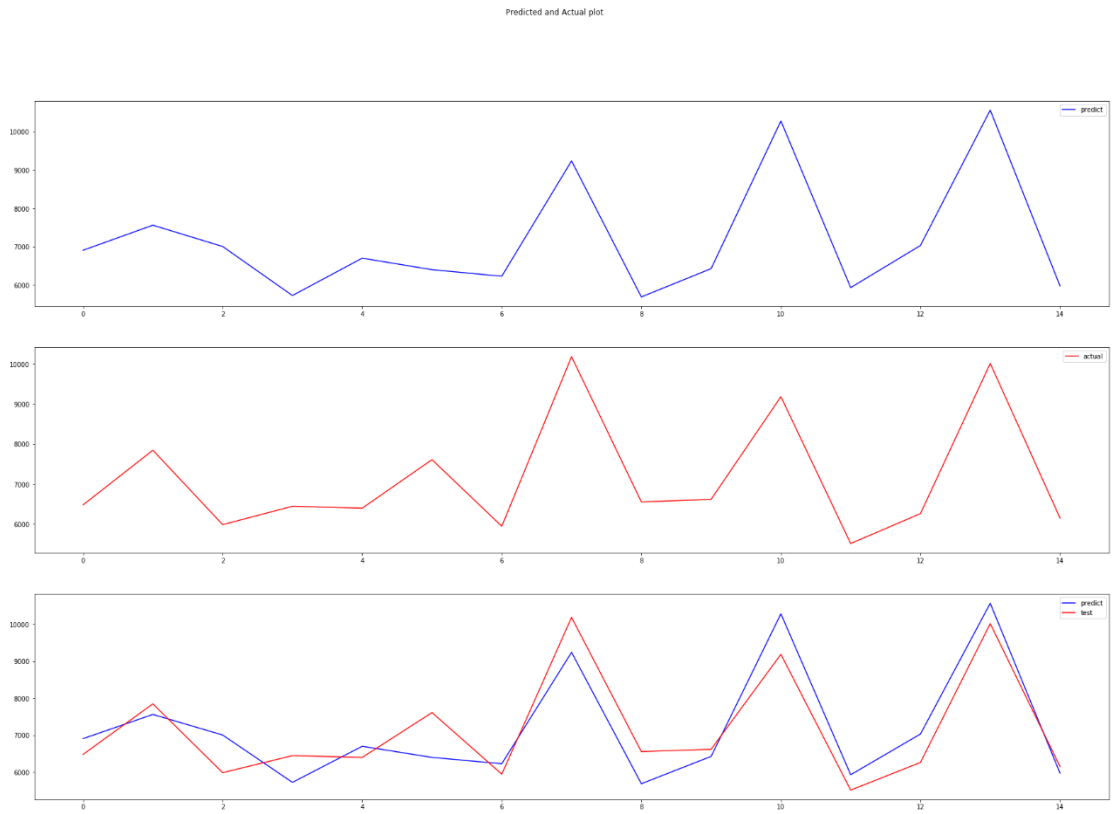
- Actual vs. Predict Visualization







- Actual vs. Prediction plot



From the visualization graphs we could see that the model is doing a relatively good job in predicting residential energy consumption.

10. Washington

1. VIF

   We calculated the VIF of all numeric independent variables in order to check for multicollinearity. The calculated VIF is as follows:

```
Commercial_Retail Price        6.849942
CLDD                           5.043081
TAVG                         264.427640
AWND                           1.274277
HTDD                         230.474229
area                       10309.172722
population                     8.577639
solar-generation               4.424532
dtype: float64
```

```
Residential_Retail Price        5.554614
CLDD                            5.035217
TAVG                          254.541202
AWND                            1.404353
HTDD                          226.138506
area                        11573.365033
population                      8.583747
solar-generation                4.984945
dtype: float64

Industrial_Retail Price         3.014607
CLDD                            7.286288
TAVG                          319.437221
AWND                            1.408602
HTDD                          274.895505
area                        11265.997181
population                      4.262276
solar-generation                4.447686
dtype: float64
```

We could observe that variable TAVG, HTDD and area have a very large VIF, indicating that there might be multicollinearity among the variables. We try to remove TAVG and area to eliminate multicollinearity, and the VIF after removal is as follows:

```
const                        3731.299307
Commercial_Retail Price         6.591829
CLDD                            2.169356
AWND                            1.216786
HTDD                            3.994829
population                      8.168222
solar-generation                4.393280
dtype: float64

const                        4904.977655
Residential_Retail Price        5.552924
CLDD                            2.151853
AWND                            1.346548
HTDD                            5.713858
population                      8.534801
solar-generation                4.952352
dtype: float64

const                        3622.501673
Industrial_Retail Price         2.401437
CLDD                            2.464148
AWND                            1.259901
HTDD                            3.989531
population                      3.783191
solar-generation                4.442072
dtype: float64
```

All VIF values are smaller than 10, indicating that there is no multicollinearity among the independent variables now.

2. Commercial Consumption

OLS Regression Results

Dep. Variable:	Commercial_Usage	R-squared:	0.674
Model:	OLS	Adj. R-squared:	0.611
Method:	Least Squares	F-statistic:	10.78
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	7.42e-09
Time:	15:38:30	Log-Likelihood:	-325.38
No. Observations:	57	AIC:	670.8
Df Residuals:	47	BIC:	691.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2270.2212	74.718	30.384	0.000	2119.908	2420.535
Commercial_Retail Price	-41.7054	116.724	-0.357	0.722	-276.524	193.113
CLDD	326.4642	79.151	4.125	0.000	167.232	485.696
AWND	-80.7075	54.306	-1.486	0.144	-189.958	28.543
HTDD	266.6917	112.467	2.371	0.022	40.438	492.945
population	64.7704	98.675	0.656	0.515	-133.739	263.280
solar-generation	40.0083	102.025	0.392	0.697	-165.239	245.256
Summer	2.8284	52.889	0.053	0.958	-103.570	109.227
Fall	-30.5637	38.507	-0.794	0.431	-108.029	46.901
Winter	110.2966	44.977	2.452	0.018	19.815	200.778

Omnibus:	1.128	Durbin-Watson:	1.960
Prob(Omnibus):	0.569	Jarque-Bera (JB):	0.814
Skew:	-0.293	Prob(JB):	0.666
Kurtosis:	3.003	Cond. No.	25.1

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.674, adjusted R-squared is 0.611. This indicates that after adjusting for the number of predictors, 61.1% of the commercial energy consumption of Washington could be explained by the independent variables.

- Feature Coefficients and Significance
    - Among the independent variables, we could see that “Retail Price” has a negative impact on the commercial energy consumption: when retail price increase by 1 dollar, the monthly commercial energy consumption of Washington decreases by about 41.7 kilo-watthours. However, this coefficient is not significant for commercial consumption in Washington, which suggests that commercial use of electricity is not that sensitive to retail prices as in other states.

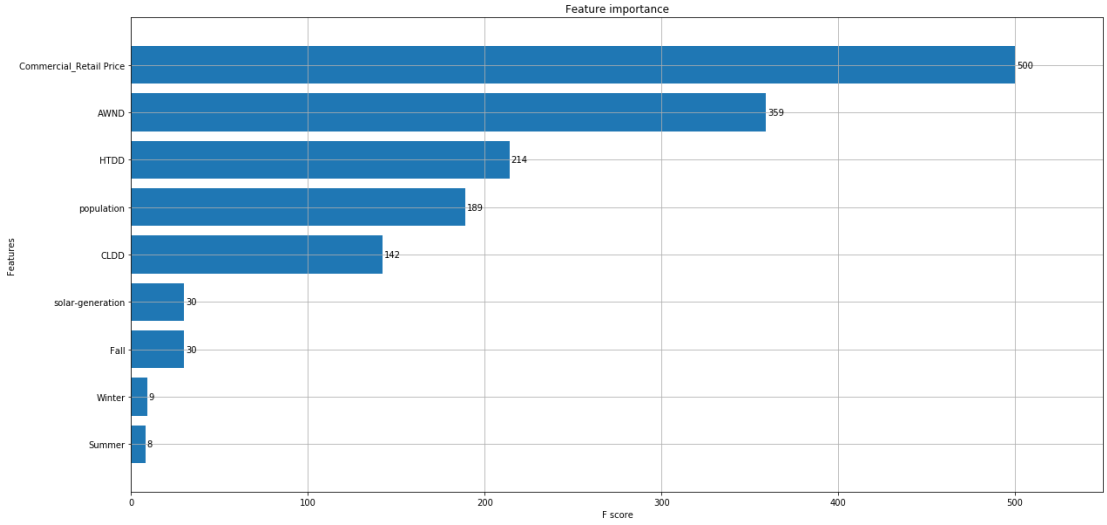
    - In this regression, we could see that CLDD and HTDD as two weather indicators are having very significant positive impact on commercial consumption: when indoors cooling days increase by 1, average monthly commercial consumption increase by 326.5 kilo-watthours, when heating

days increase by 1, average monthly commercial consumption increase by 566.7 kilo-watthours. This could because commercial consumption is taken up a lot by the indoor air conditioning in Washington.

- Winter as a seasonal dummy variable is being significant at 5% level of confidence. This indicates that, compared with spring, the commercial energy consumption in winter is about 110.3 kilo-watthours higher per month.

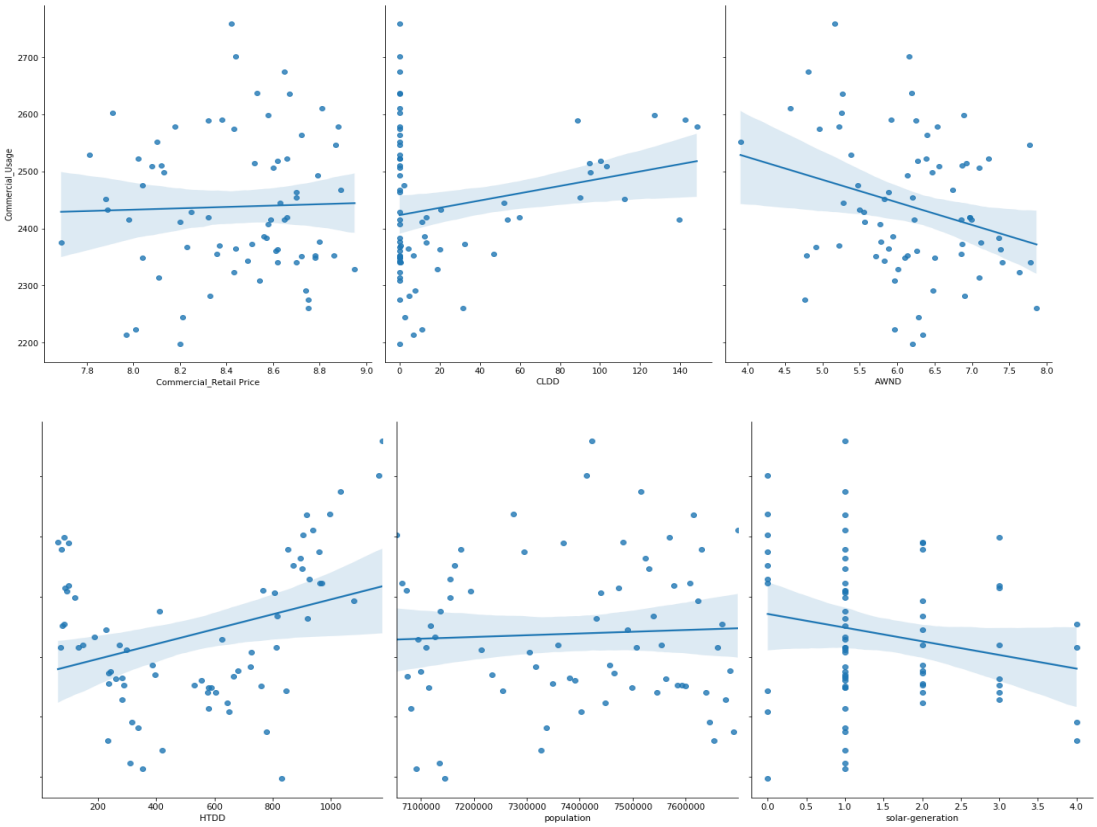
- MAE, MSE, RMSE
  - Mean Absolute Error: 66.366667265509
  - Mean Squared Error: 5850.257940881382
  - Root Mean Squared Error: 76.48697889759656

- Feature Importance

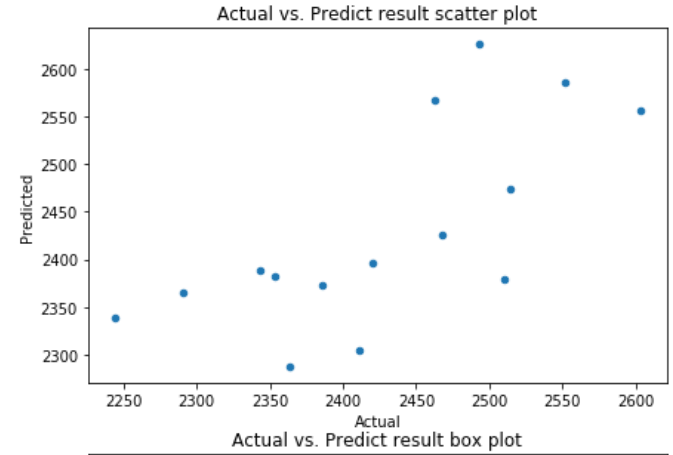


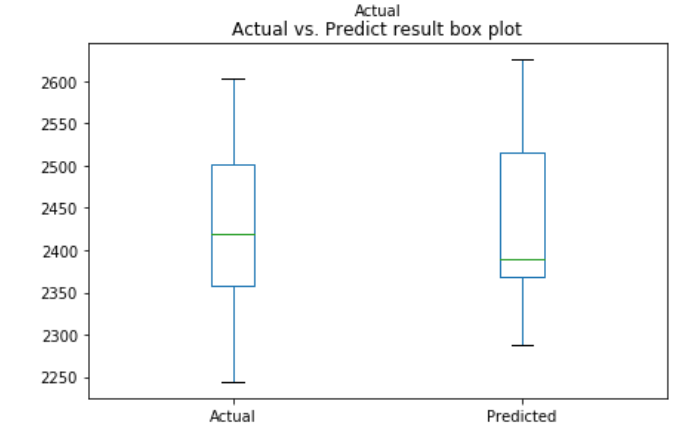
In all the features, top four features that contribute the most to the explanation of energy consumption are: Retail Price of electricity, AWND, HTDD and Population.

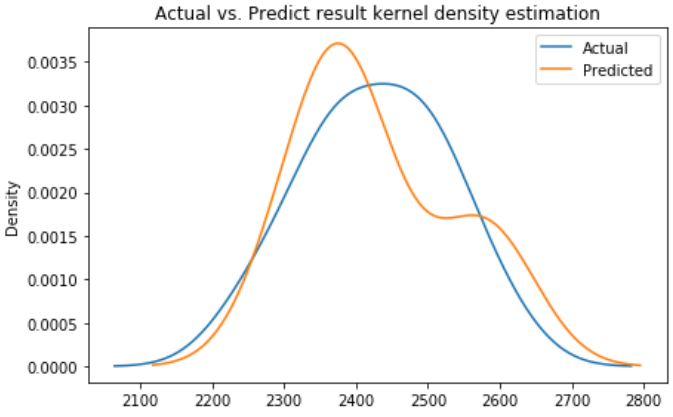
- Correlation Analysis



- Actual vs. Predict Visualization







- Actual vs. Prediction plot

3. Industrial Consumption

OLS Regression Results

Dep. Variable:	Industrial_Usage	R-squared:	0.773
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	17.83
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	2.22e-12
Time:	15:38:34	Log-Likelihood:	-319.43
No. Observations:	57	AIC:	658.9
Df Residuals:	47	BIC:	679.3
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2339.1307	65.401	35.766	0.000	2207.560	2470.701
Industrial_Retail Price	-109.8164	87.372	-1.257	0.215	-285.587	65.954
CLDD	108.7562	82.156	1.324	0.192	-56.521	274.033
AWND	-109.2767	48.687	-2.244	0.030	-207.222	-11.331
HTDD	5.3478	108.051	0.049	0.961	-212.022	222.718
population	-304.9638	61.697	-4.943	0.000	-429.082	-180.845
solar-generation	114.0688	97.832	1.166	0.250	-82.743	310.881
Summer	51.4051	46.167	1.113	0.271	-41.471	144.282
Fall	-17.0233	36.336	-0.468	0.642	-90.121	56.075
Winter	-23.2318	40.930	-0.568	0.573	-105.573	59.109

Omnibus:	0.182	Durbin-Watson:	2.161
Prob(Omnibus):	0.913	Jarque-Bera (JB):	0.193
Skew:	-0.121	Prob(JB):	0.908
Kurtosis:	2.848	Cond. No.	25.8

- R-squared, Adjusted R-squared
    - The value of R-squared is 0.773, adjusted R-squared is 0.730. This indicates that after adjusting for the number of predictors, 73.0% of the industrial energy consumption of Washington could be explained by the independent variables.

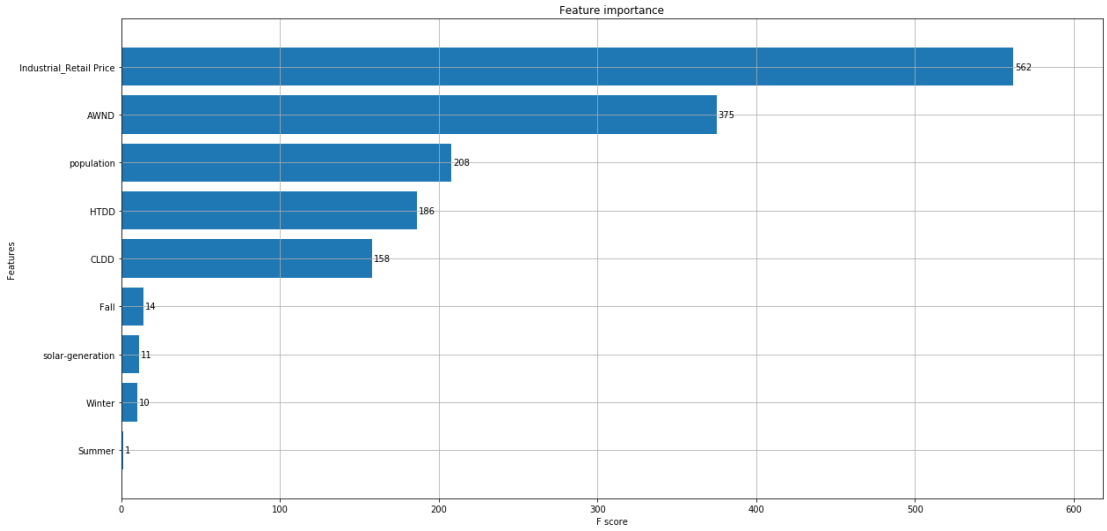
- Feature Coefficients and Significance
    - Among the independent variables, we could see that “Retail Price” has a negative impact on the industrial energy consumption: when retail price increase by 1 dollar, the monthly industrial energy consumption of Washington decreases by about 109.81 kilo-watthours. However, this coefficient is not significant for industrial consumption in Washington, which suggests that industrial use of electricity is not that sensitive to retail prices as in other states.

    - “AWND” is also a significant feature with a p-value of -2.244. This indicates that when average wind speed increase by 1 m/s, average industrial energy consumption decreases by 109.27 kilo-watthours.

- MAE, MSE, RMSE
    - Mean Absolute Error: 64.73563256207665
    - Mean Squared Error: 5678.970929892141

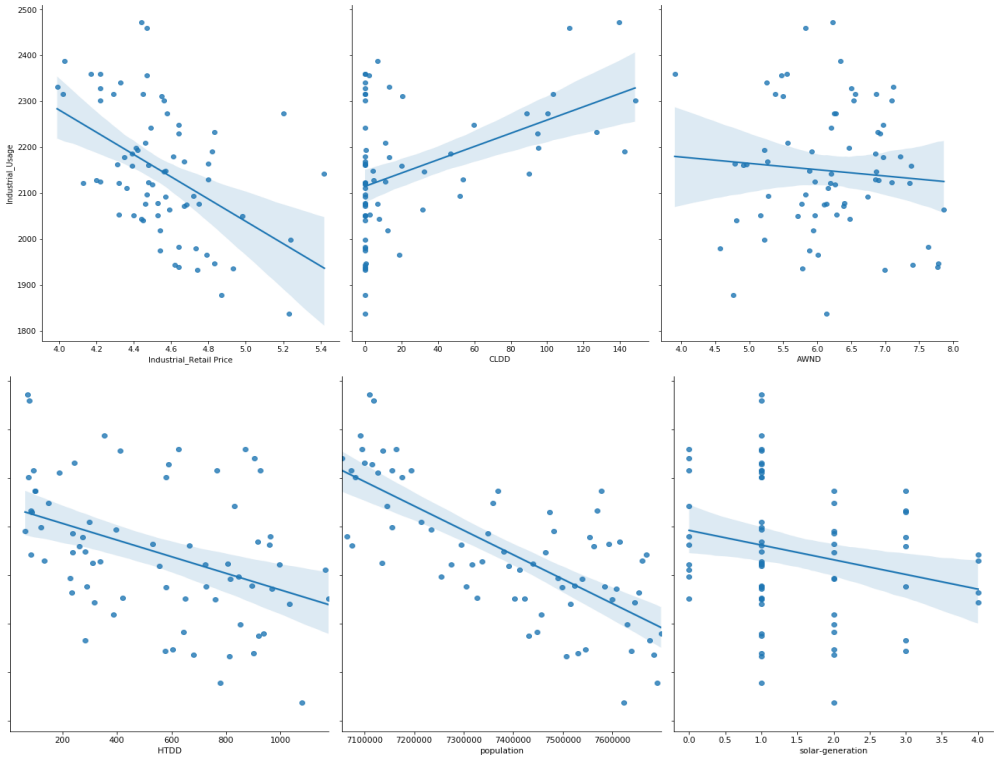
- Root Mean Squared Error: 75.35894724511576

- Feature Importance

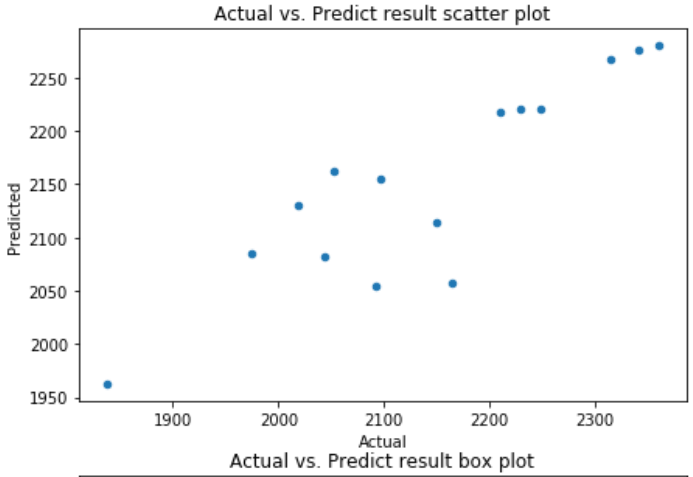


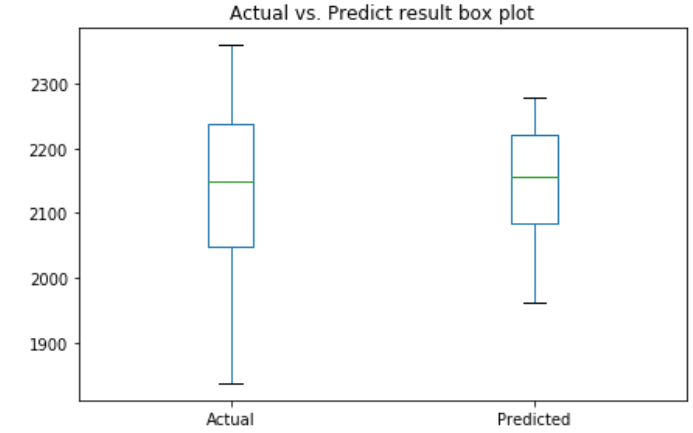
In all the features, top four features that contribute the most to the explanation of energy consumption are: Retail Price of electricity, AWND, Population and HTDD. Here, CLDD does not matter that much in industrial consumption in Washington. This could because Washington is in the northern part of the country and therefore more sensitive to heating days but not cooling days.

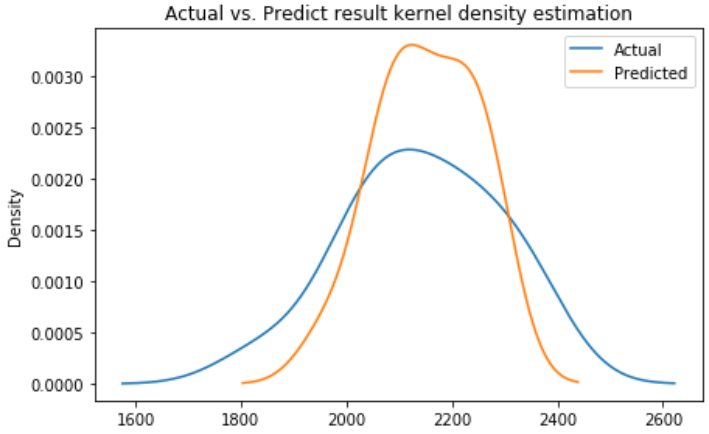
- Correlation Analysis



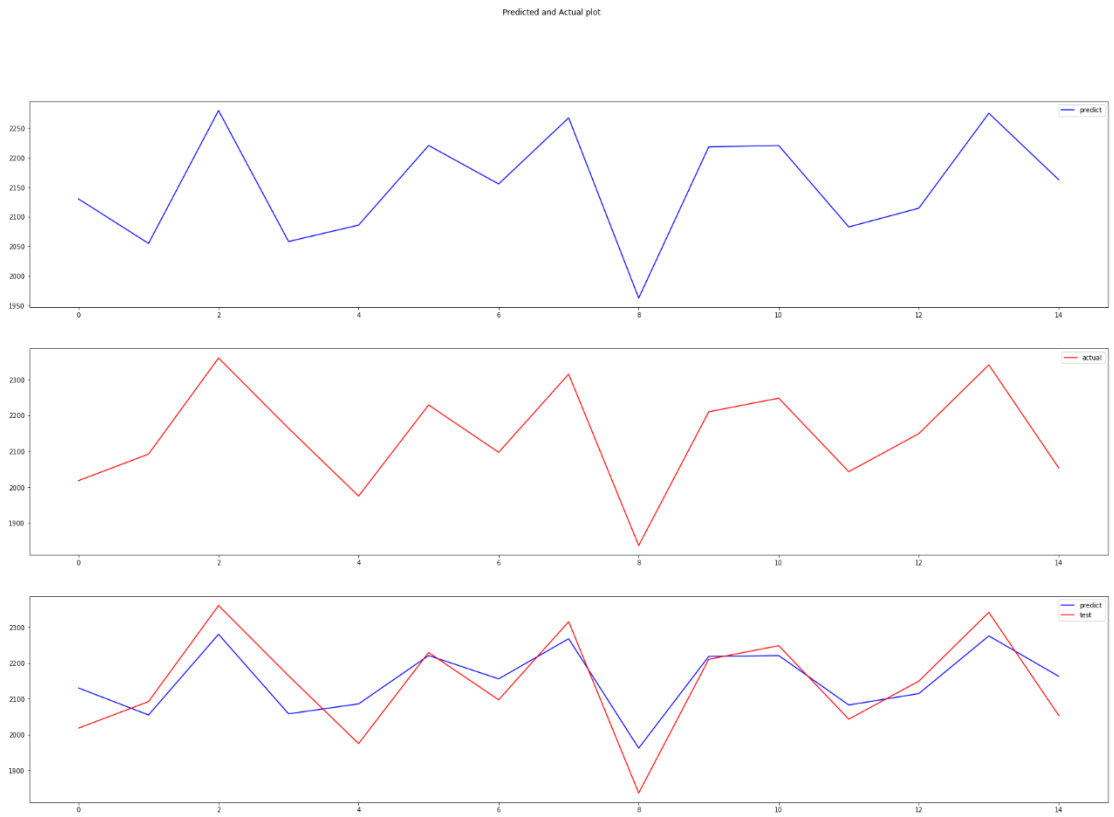
- Actual vs. Predict Visualization







- Actual vs. Prediction plot



From the visualization graphs we could see that the model is doing a relatively good job in predicting industrial energy consumption.

4. Residential Consumption

OLS Regression Results

Dep. Variable:	Residential_Usage	R-squared:	0.961
Model:	OLS	Adj. R-squared:	0.954
Method:	Least Squares	F-statistic:	129.7
Date:	Tue, 21 Jul 2020	Prob (F-statistic):	4.23e-30
Time:	15:38:37	Log-Likelihood:	-359.08
No. Observations:	57	AIC:	738.2
Df Residuals:	47	BIC:	758.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1827.5773	154.920	11.797	0.000	1515.918	2139.237
Residential_Retail Price	161.0175	194.596	0.827	0.412	-230.460	552.495
CLDD	779.4656	142.031	5.488	0.000	493.736	1065.195
AWND	-322.2110	102.915	-3.131	0.003	-529.250	-115.172
HTDD	2467.6331	221.080	11.162	0.000	2022.878	2912.388
population	-39.2457	180.234	-0.218	0.829	-401.831	323.339
solar-generation	198.2378	188.361	1.052	0.298	-180.695	577.171
Summer	-19.0670	93.233	-0.205	0.839	-206.627	168.493
Fall	-342.2931	68.932	-4.966	0.000	-480.966	-203.620
Winter	138.2188	81.688	1.692	0.097	-26.116	302.554

Omnibus:	3.980	Durbin-Watson:	2.114
Prob(Omnibus):	0.137	Jarque-Bera (JB):	3.693
Skew:	0.239	Prob(JB):	0.158
Kurtosis:	4.152	Cond. No.	29.6

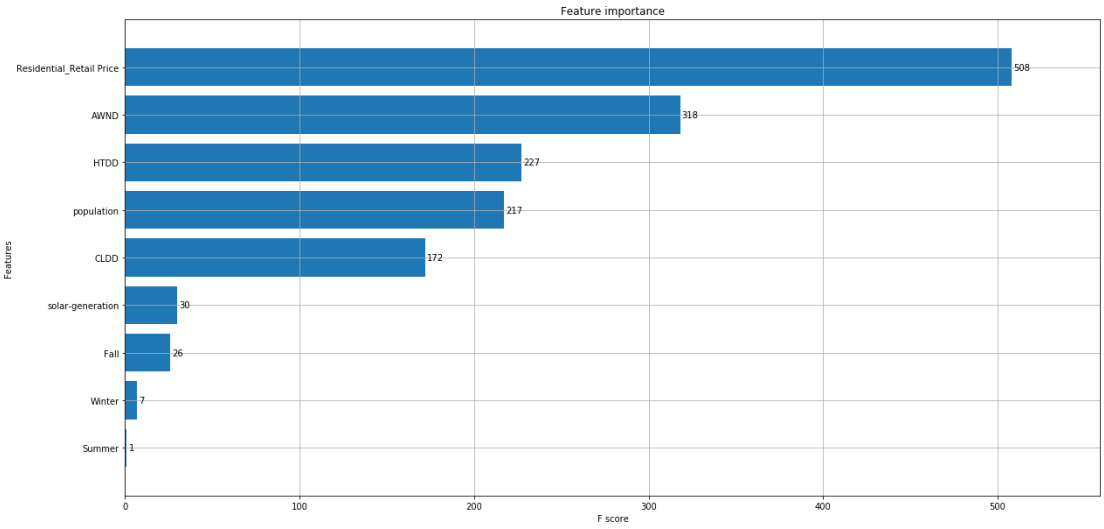
- R-squared, Adjusted R-squared
    - The value of R-squared is 0.961, adjusted R-squared is 0.954. This indicates that after adjusting for the number of predictors, 95.4% of the residential energy consumption of Washington could be explained by the independent variables.

- Feature Coefficients and Significance
    - In this regression, we could see that CLDD and HTDD as two weather indicators are having very significant positive impact on residential consumption: when indoors cooling days increase by 1, average monthly commercial consumption increase by 779 kilo-watthours, when heating days increase by 1, average monthly commercial consumption increase by 2467 kilo-watthours. This could because residential consumption is taken up a lot by the indoor air conditioning in Washington.

      Noticeably, heating days has a greater impact on residential energy consumption compared with cooling days, this could because of Washington's geographic and climate characteristics, that average temperature is pretty low and cooling is not as needed in summer as heating is needed in winter.

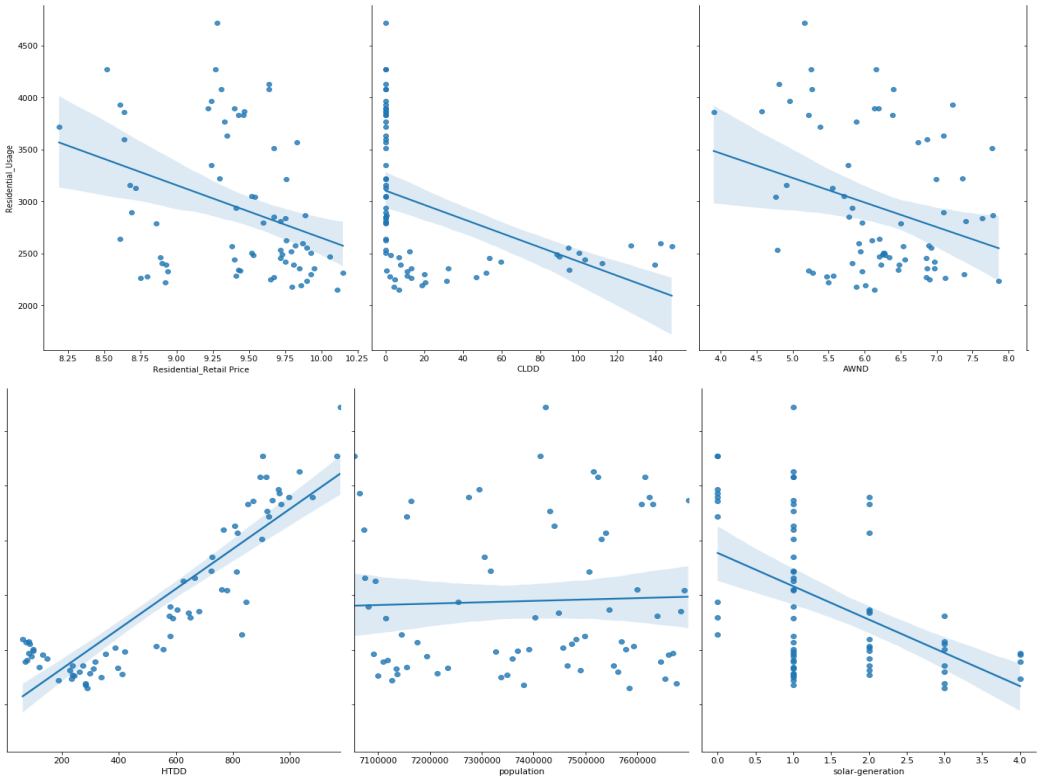
- MAE, MSE, RMSE
    - Mean Absolute Error: 187.06400464384205
    - Mean Squared Error: 48214.10722356022
    - Root Mean Squared Error: 219.5771099717824

- Feature Importance

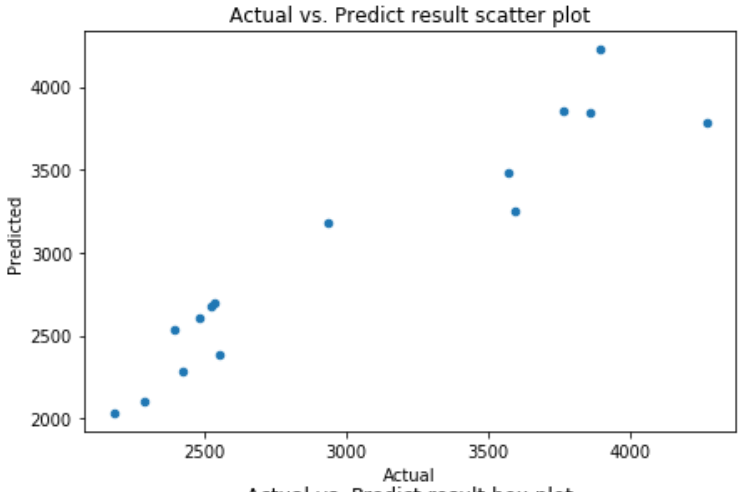


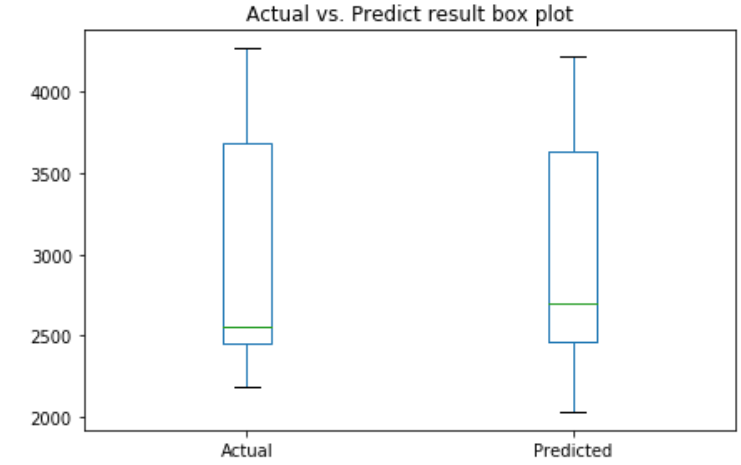
In all the features, top four features that contribute the most to the explanation of energy consumption are: Retail Price of electricity, AWND, HTDD and Population. Here, CLDD does not matter that much in commercial consumption in Washington. This could because Washington is in the northern part of the country and therefore more sensitive to heating days but not cooling days.

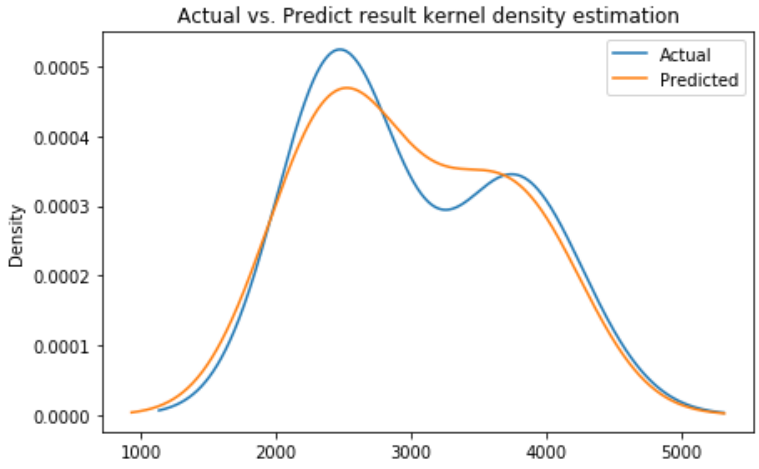
- Correlation Analysis



- Actual vs. Predict Visualization







- Actual vs. Prediction plot

From the visualization graphs we could see that the model is doing a relatively good job in predicting residential energy consumption.

5.3.3.2 Artificial Neural Network(ANN)

In the previous section, we had applied the traditional linear regression model to forecast state-wise electricity consumption. Even though the predicted values are mostly similar to the actual values, the MAEs and RMSEs are quite large that are able to be improved. Therefore, we decided to apply deep learning techniques to improve our model performance.

The method we try out to apply to our dataset is Artificial Neural Network(ANN). Artificial neural networks are one of the main tools used in machine learning. They are brain-inspired systems that are intended to replicate the way that we humans learn (Dormehl, 2019). Since we assume that there might be hidden patterns or non-linear patterns that cannot be shown by regression analysis, we determined to apply ANN to our dataset to see if there are hidden patterns that are able to be explained by deep learning models.

We chose five states' data to apply ANN analysis in which their regression analysis performance is worse and possible to be improved. Moreover, we adjusted the ANN model by adding month data into features since including time data into a deep learning model is a common approach(which is difficult to include in the regression model). Since the training data might differ when doing random train-test splits in both models, the actual values shown below might differ when showing plots.

In the following section, we mainly compared our ANN performance with our regression analysis performance to see the difference.

1. **Pennsylvania**
   - Residential
     i. MAE, MSE, RMSE Comparison
        - The performance of regression analysis:

```
Mean Absolute Error: 202.7552706066282
Mean Squared Error: 63779.46390997826
Root Mean Squared Error: 252.54596395503583
```

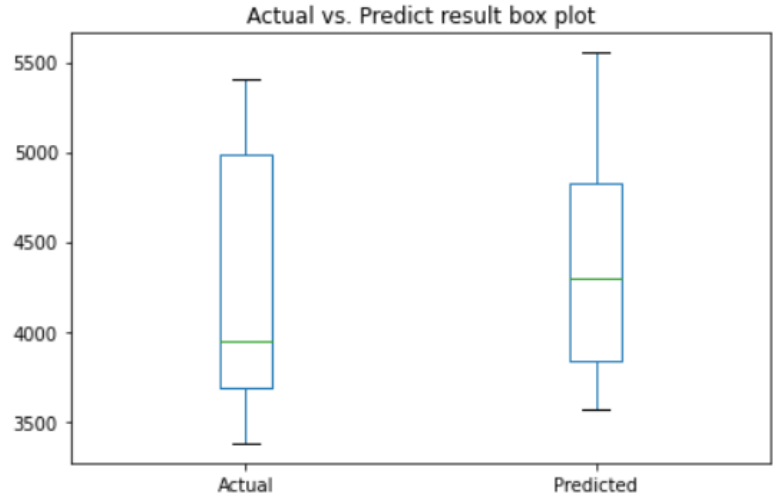
        - The performance of ANN:

```
Mean Absolute Error: 151.51472981770834
Mean Squared Error: 36075.00347748995
Root Mean Squared Error: 189.93420828668528
```

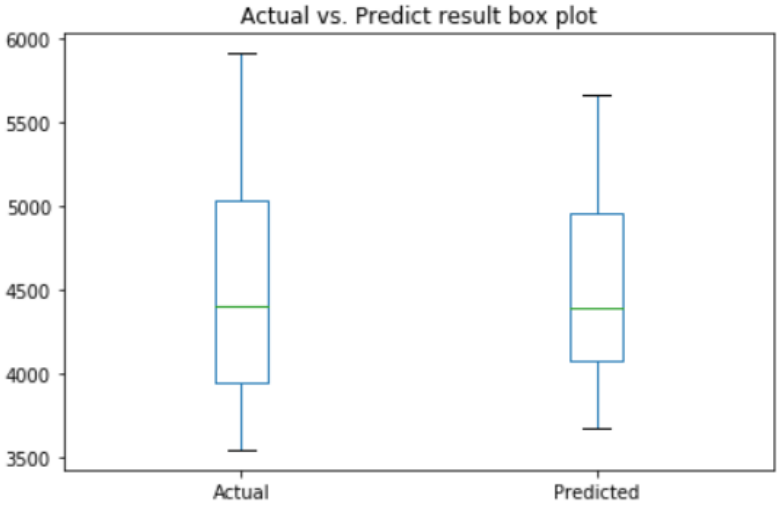
Compare to regression analysis, the MAE, MSE, and RMSE drop a lot, which means that the model performs better than the regression model. In other words, the relationship between residential electricity consumption and other features in Pennsylvania is not likely to be linear, there are likely to have hidden patterns between the features and electricity consumption.

ii. Box plot Comparison

- Regression Analysis box plot



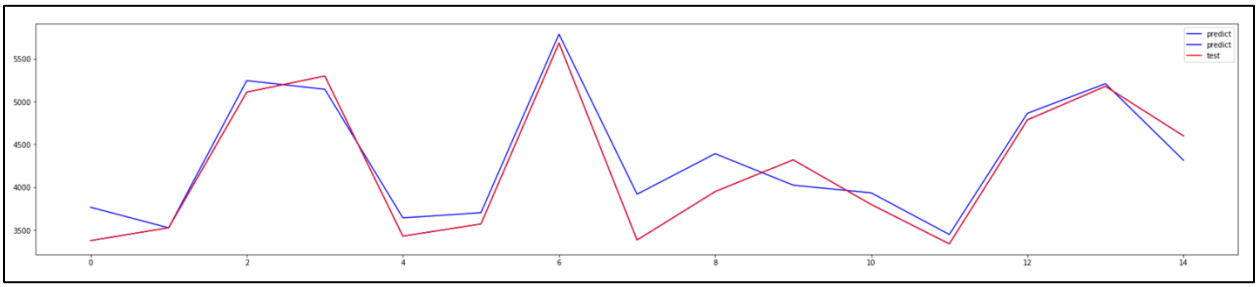
- ANN box plot



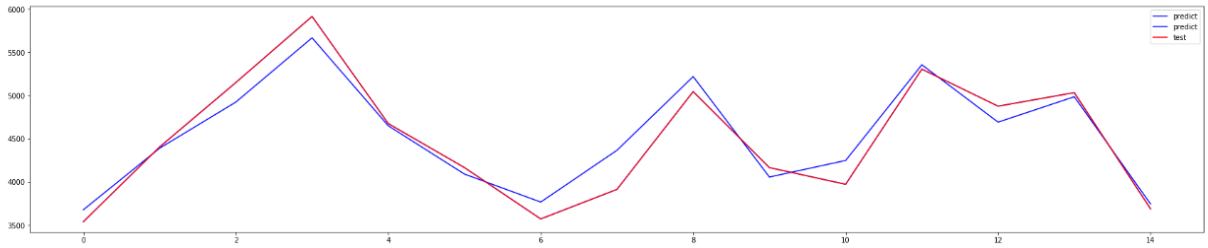
From the box plots, we can see that ANN model predict much better than regression model. The bias and variance are both smaller.

iii. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, the ANN model also shows better fit than the regression model.

In conclusion, ANN model performs better than regression model while predicting residential electricity consumption in Pennsylvania.

- Industrial
  i. MAE, MSE, RMSE Comparison
    - The performance of regression analysis:

```
Mean Absolute Error: 171.630089969333
Mean Squared Error: 35464.04925772611
Root Mean Squared Error: 188.31900928405
```

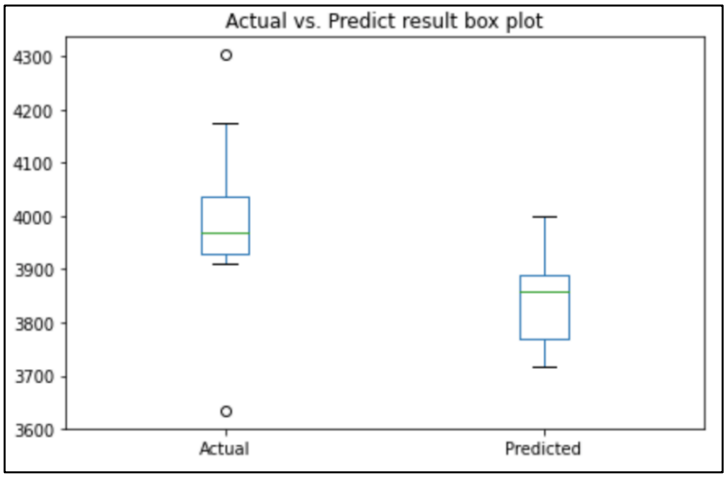
    - The performance of ANN:

```
Mean Absolute Error: 187.14890950520834
Mean Squared Error: 50888.43087854783
Root Mean Squared Error: 225.5846423818515
```

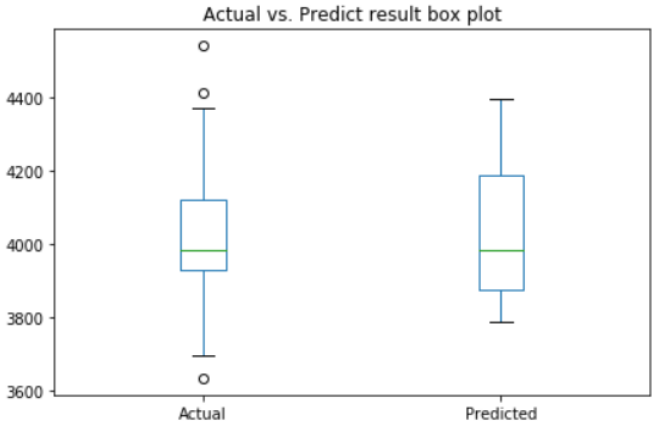
Compare to regression analysis, the MAE, MSE, and RMSE increase a lot, which means that the model performs worse than the regression model. In other words, the relationship between industrial electricity consumption and other features in Pennsylvania is likely to be linear.

ii. Box plot Comparison

- Regression Analysis box plot



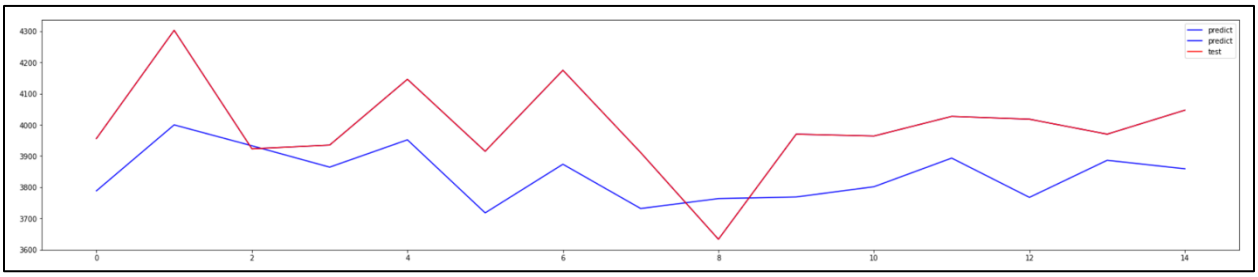
- ANN box plot



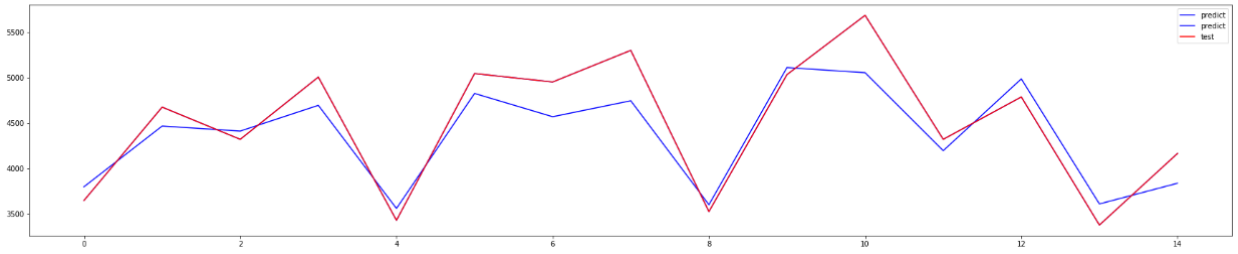
From the box plots, we can see that ANN model predict a high variance but low bias model than the regression model.

iii. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, they show that both models do not fit the data well, but ANN model fits better.

By comparing the performance of the two models, we can conclude that ANN model performs better than regression model when predicting industrial electricity consumption in Pennsylvania.

- Commercial

iv. MAE, MSE, RMSE Comparison

- The performance of regression analysis:

```
Mean Absolute Error: 215.5343495979653
Mean Squared Error: 92342.39217748676
Root Mean Squared Error: 303.87891038617136
```

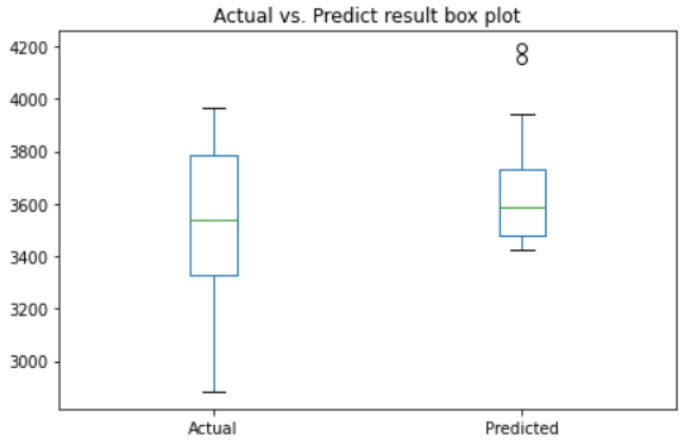
- The performance of ANN:

```
Mean Absolute Error: 136.94098307291668
Mean Squared Error: 25488.125499065718
Root Mean Squared Error: 159.65000939262646
```

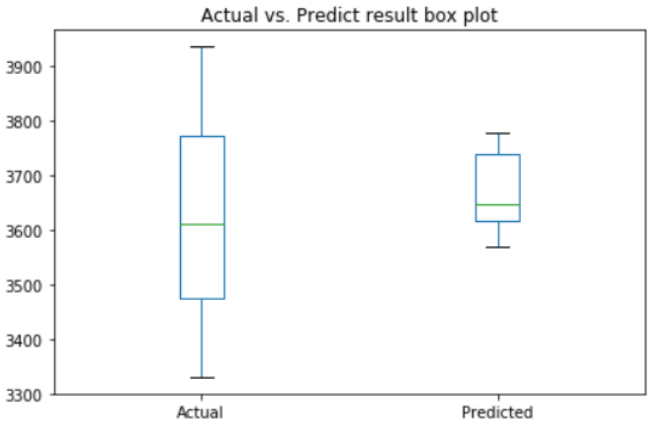
Compare to regression analysis, the MAE, MSE, and RMSE drop a lot, which means that the model performs better than the regression model. In other words, the relationship between commercial electricity consumption and other features in Pennsylvania is not likely to be linear, there are likely to have hidden patterns between the features and electricity consumption.

v. Box plot Comparison

- Regression Analysis box plot



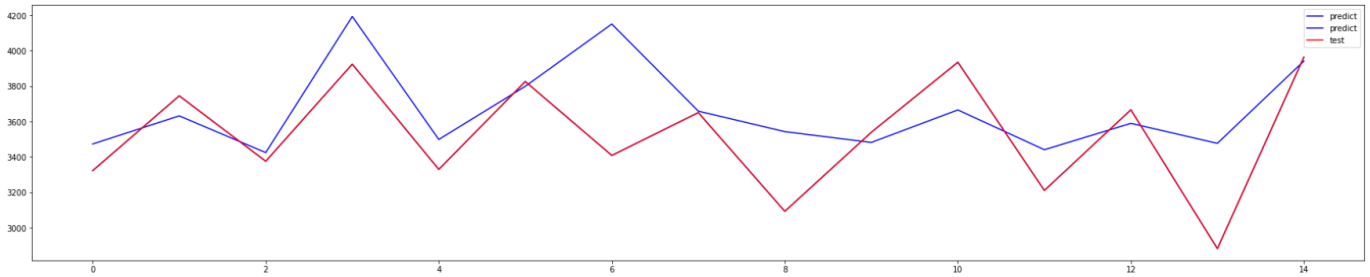
- ANN box plot



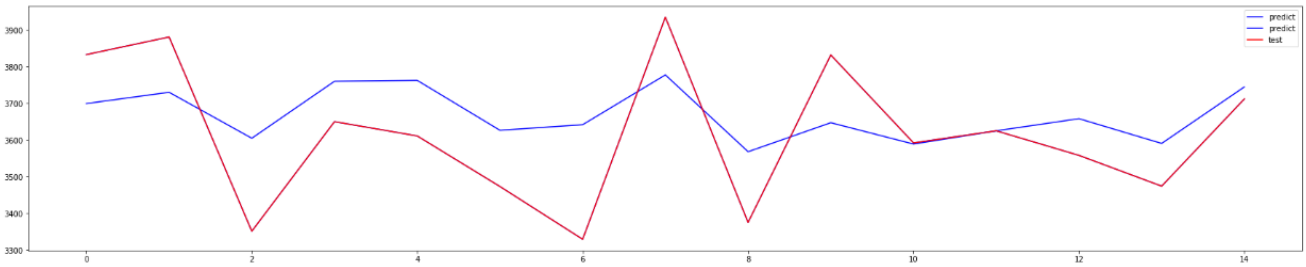
From the box plots, we can see that both models only predict in a small value range compare to the actual values, which means that both models are unable to predict the variance of actual values.

vi. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, they show that both models do not fit the data well.

By comparing the performance of the two models, we can conclude that both regression model and ANN do not fit the data well. Even though ANN model has smaller MAE and RMSE, the plots show that it poor performance. The reason might because of we did not find out high-related features to predict commercial electricity consumption in Pennsylvania, therefore the bias is too large for us to find an appropriate model to make prediction.

2. California

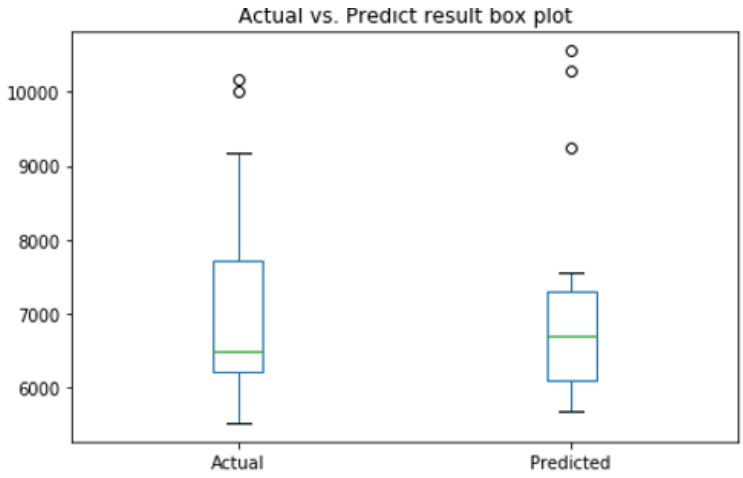
- Residential
  - i. MAE, MSE, RMSE Comparison
    - The performance of regression analysis:

```
Performance Evaluation
Mean Absolute Error: 616.7428590053264
Mean Squared Error: 495157.3207022466
Root Mean Squared Error: 703.674158046355
```

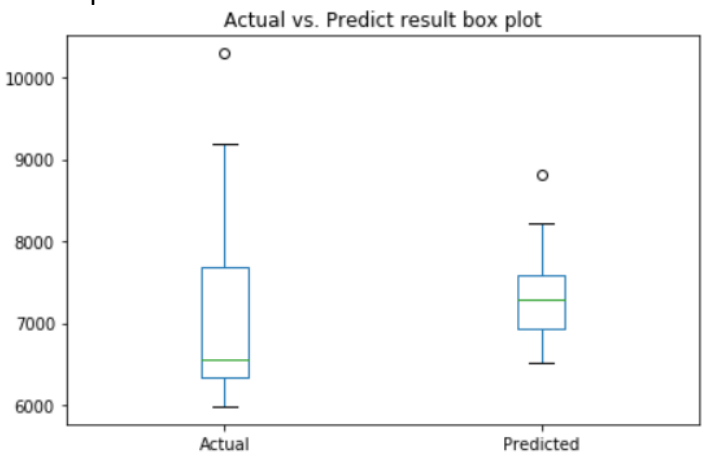
    - The performance of ANN:

```
Mean Absolute Error: 647.88466796875
Mean Squared Error: 537374.322924757
Root Mean Squared Error: 733.0581988660634
```

      Both models have similar MAE, MSE and RMSE, and the values are large, which means that both models do not fit the data well.
  - ii. Box plot Comparison
    - Regression Analysis box plot



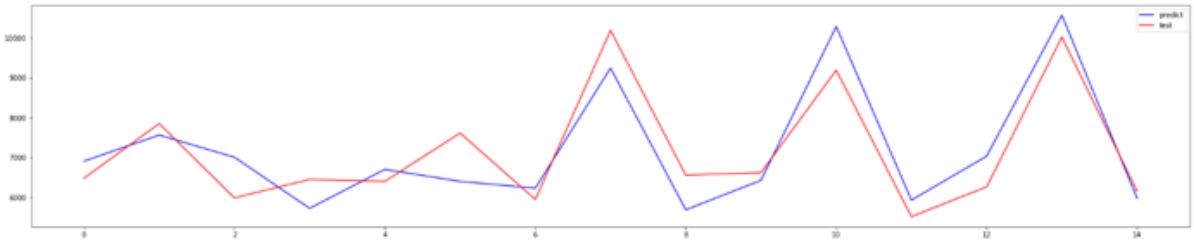
- ANN box plot



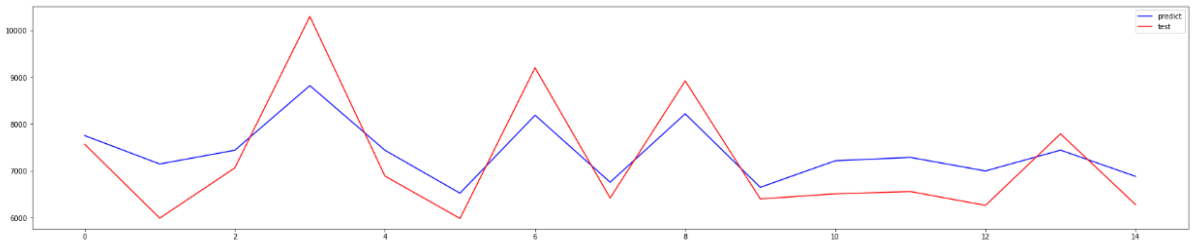
From the box plots, we can see that regression model predict better than ANN model. However, the regression model is still unable to predict the variance of the residential electricity consumption in Pennsylvania.

iii. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, the regression model also shows better fit than the ANN model.

In conclusion, regression model performs better than regression model while predicting residential electricity consumption in Pennsylvania. However, the MSE and MAE are both so large. Therefore, it is better to find a more robust model to predict residential electricity consumption in California.

- Industrial
  - i. MAE, MSE, RMSE Comparison
    - The performance of regression analysis:

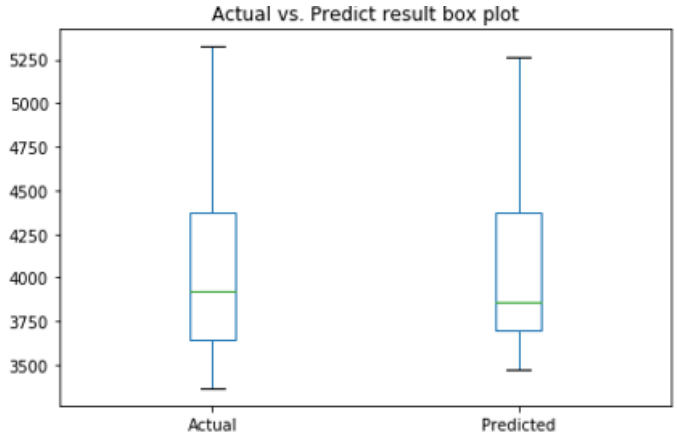
      ```
      Performance Evaluation
      Mean Absolute Error: 197.66721769206285
      Mean Squared Error: 56214.810934014524
      Root Mean Squared Error: 237.09662784192972
      ```

    - The performance of ANN:

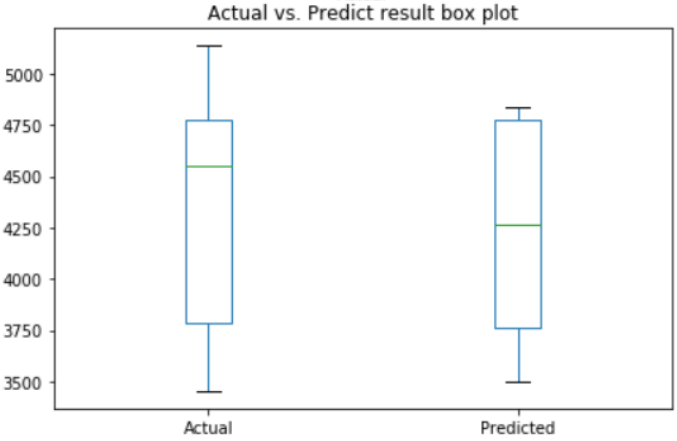
      ```
      Mean Absolute Error: 173.302099609375
      Mean Squared Error: 48948.00929119984
      Root Mean Squared Error: 221.24197000388475
      ```

      Compare to regression analysis, the MAE, MSE, and RMSE improve a bit, which means that the model performs better than the regression model. In other words, the relationship between industrial electricity consumption and other features in California is likely not to be linear. There might be hidden patterns.

  - ii. Box plot Comparison
    - Regression Analysis box plot



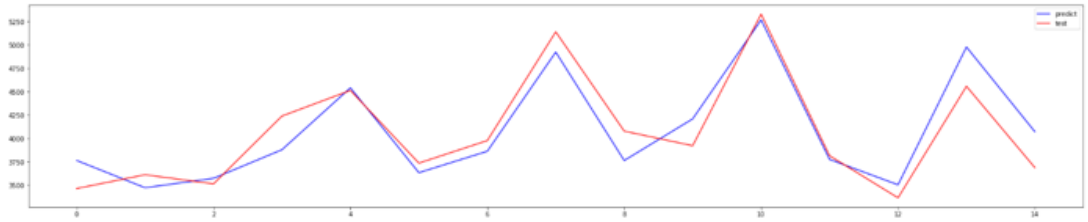
- ANN box plot



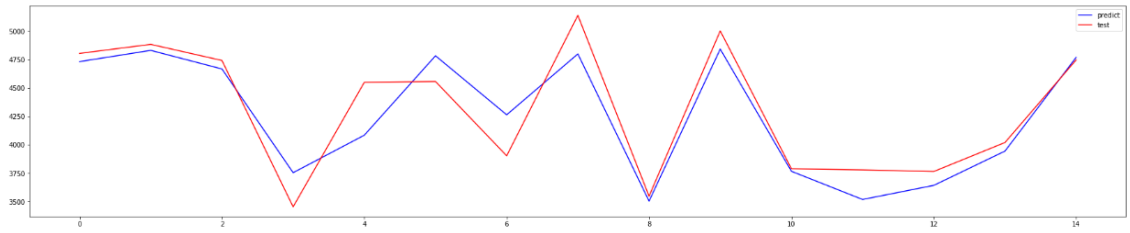
Both box plots show quiet good prediction, but the regression model predict the shape and variance better than ANN model.

iii. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, they show that both models do not fit the data well, but the regression model's line plot fits much better.

By comparing the performance of the two models, we can conclude that regression model is a better model to predict industrial electricity consumption in California. Even though ANN model performs smaller MSE and MAE, it might because that the data variate a lot in this dataset and ANN model is more robust to extreme values compare to regression model.

- Commercial
  - i. MAE, MSE, RMSE Comparison
    - The performance of regression analysis:

```
Performance Evaluation
Mean Absolute Error: 476.50229160348573
Mean Squared Error: 328613.89826780366
Root Mean Squared Error: 573.2485484218897
```

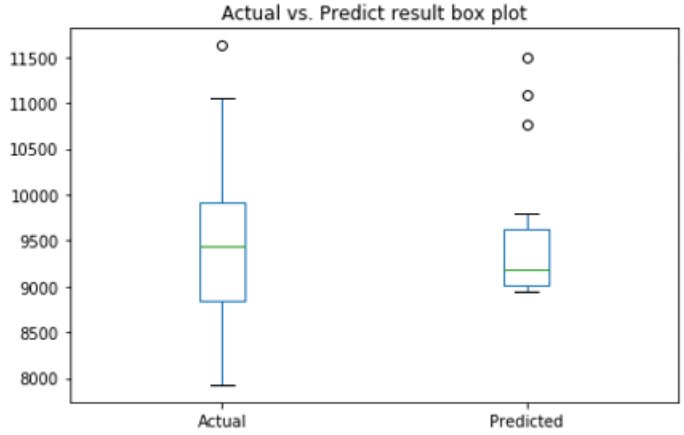
    - The performance of ANN:

```
Performance Evaluation
Mean Absolute Error: 621.9111979166667
Mean Squared Error: 571245.6222910563
Root Mean Squared Error: 755.8079268511652
```

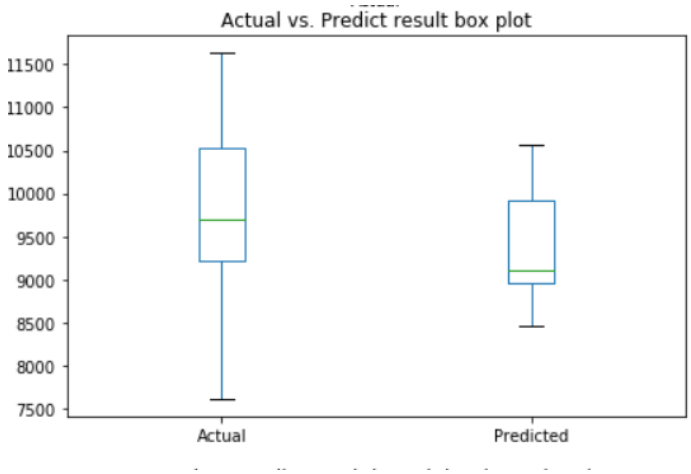
    - Compare to regression analysis, the MAE, MSE, and RMSE increase a lot, which means that the model performs worse than the regression model. In other words, the relationship between commercial electricity consumption and other features in Pennsylvania is likely to be linear than having hidden patterns. However, the values are larger. It is better to find a more appropriate model to fit the data.

  - ii. Box plot Comparison

- Regression Analysis box plot



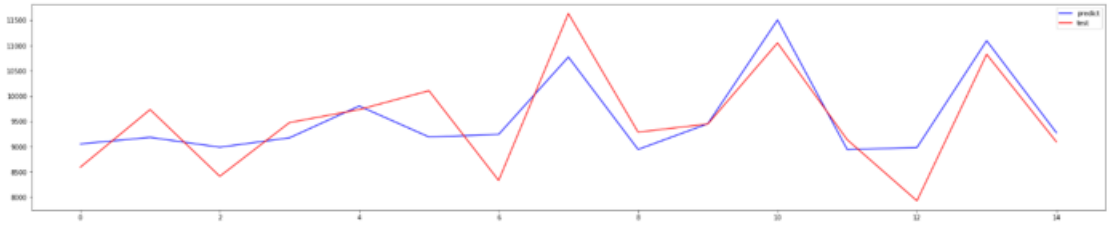
- ANN box plot



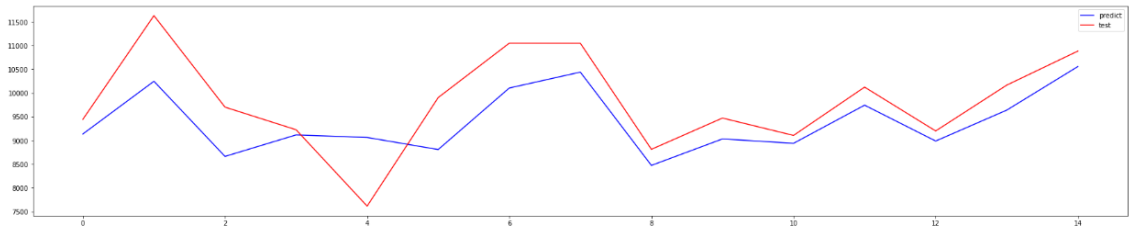
From the box plots, we can see that both models perform bad on predicting commercial electricity consumption in California.

iii. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, they show that both models do not fit the data well.

By comparing the performance of the two models, we can conclude that both regression model and ANN do not fit the data well. Even though regression model has smaller MAE and RMSE, the plots show that it poor performance. The reason might because of we did not find out high-related features to predict commercial electricity consumption in Pennsylvania, therefore the bias is too large for us to find appropriate model to make prediction.

3. Minnesota

- Residential
  - i. MAE, MSE, RMSE Comparison
    - The performance of regression analysis:

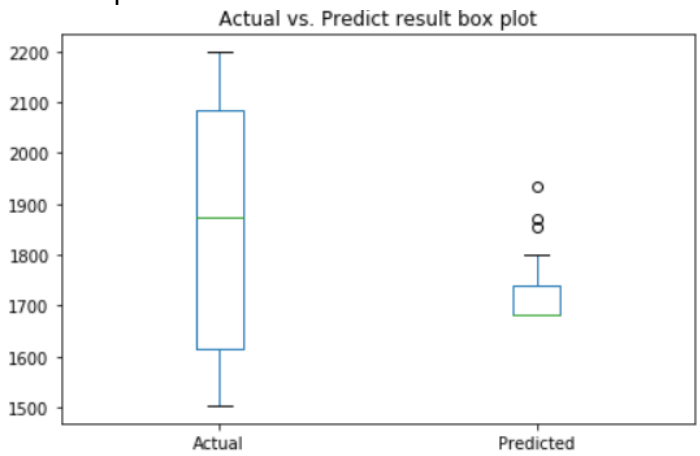
```
Mean Absolute Error: 69.40731341290339
Mean Squared Error: 6796.178504100094
Root Mean Squared Error: 82.43893803355363
```

    - The performance of ANN:

```
Mean Absolute Error: 202.852783203125
Mean Squared Error: 71101.74463890592
Root Mean Squared Error: 266.6491039529402
```

      The regression model performs much better than the ANN model.
  - ii. Box plot Comparison
    - Regression Analysis box plot

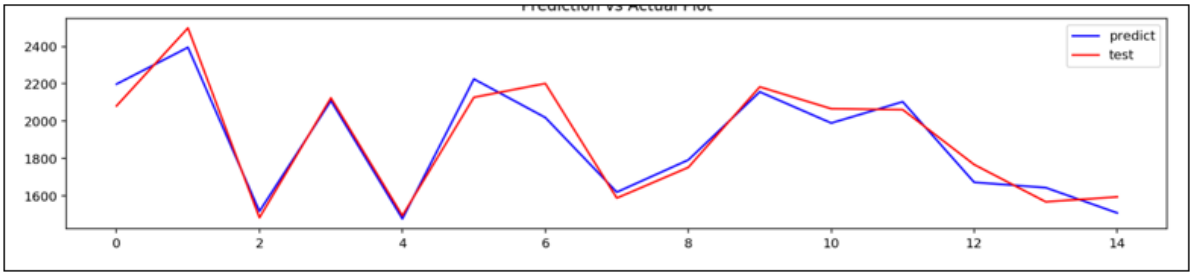
- ANN box plot



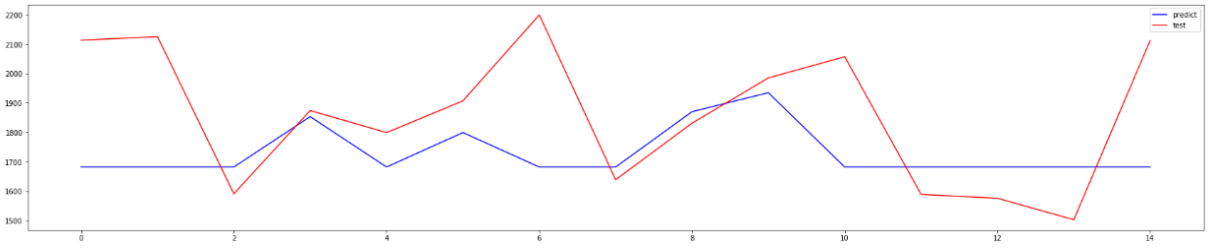
From the box plots, we can see that regression model predict much better than ANN model.

iii. Line plot Comparison

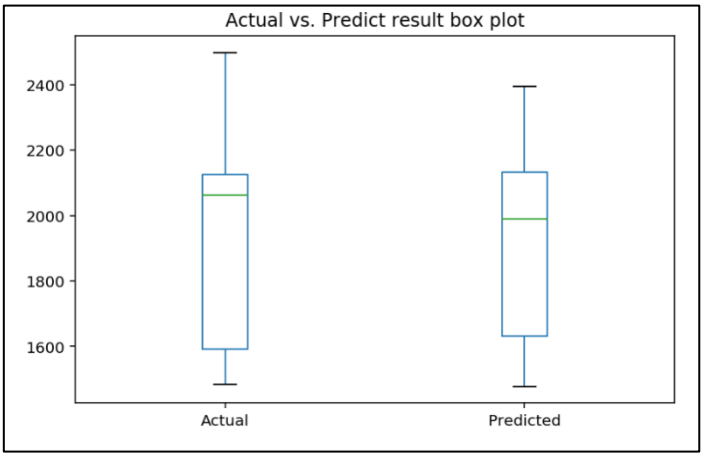
- Regression Analysis Line plot



- ANN Line plot



From the line plots, the regression model also shows better fit than the ANN model.



In conclusion, regression model performs better than regression model while predicting residential electricity consumption in Minnesota. Therefore, regression model is enough to predict the residential electricity consumption in Minnesota.

- Industrial
  i. MAE, MSE, RMSE Comparison
     - The performance of regression analysis:

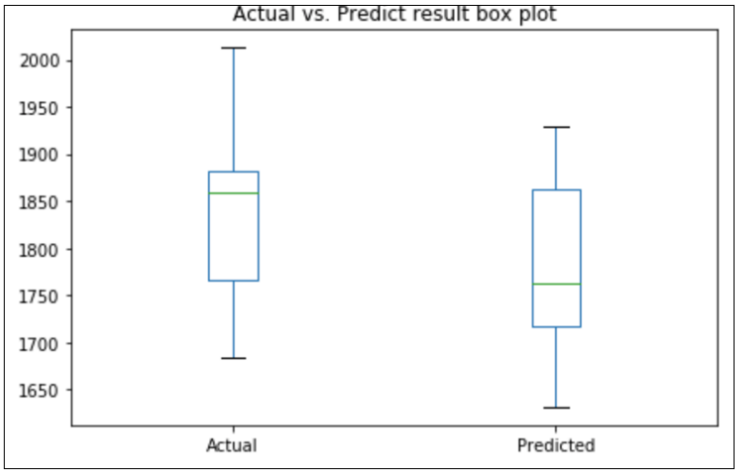
```
Performance Evaluation
Mean Absolute Error: 67.61194441108441
Mean Squared Error: 7238.424859798387
Root Mean Squared Error: 85.07893311389364
```

     - The performance of ANN:

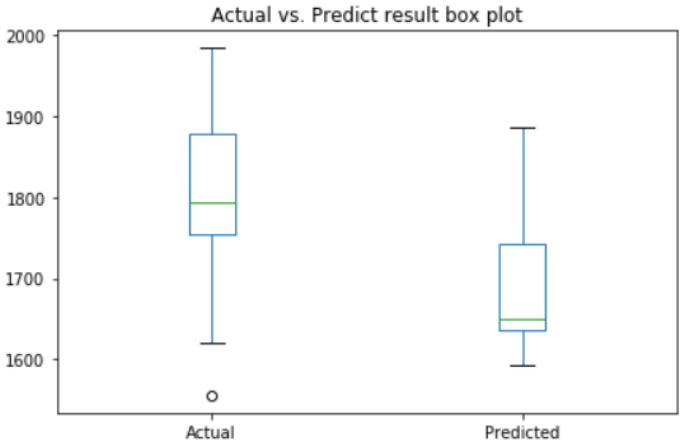
```
Mean Absolute Error: 128.4289306640625
Mean Squared Error: 22717.8308832854
Root Mean Squared Error: 150.72435398198064
```

       The regression model performs much better than ANN model.

  ii. Box plot Comparison
     - Regression Analysis box plot



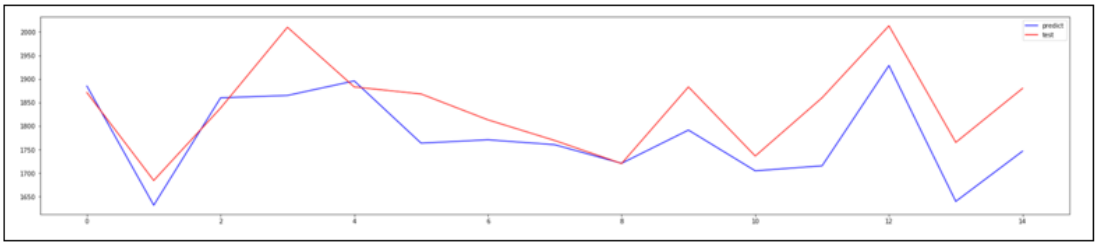
     - ANN box plot



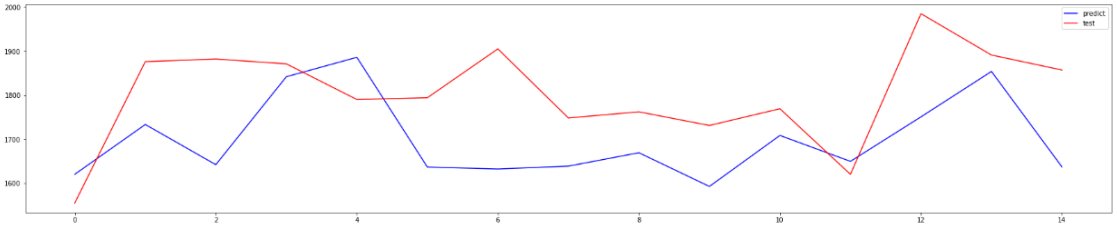
Both box plots do not show quiet good prediction.

iii. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, they show that both models do not fit the data well, but the regression model's line plot fits much better.

By comparing the performance of the two models, we can conclude that regression model is a better model to predict industrial electricity consumption in Minnesota. However, from the plots shown above, it is possible that there are more models to try to get better performance.

- Commercial
  - iv. MAE, MSE, RMSE Comparison
    - The performance of regression analysis:

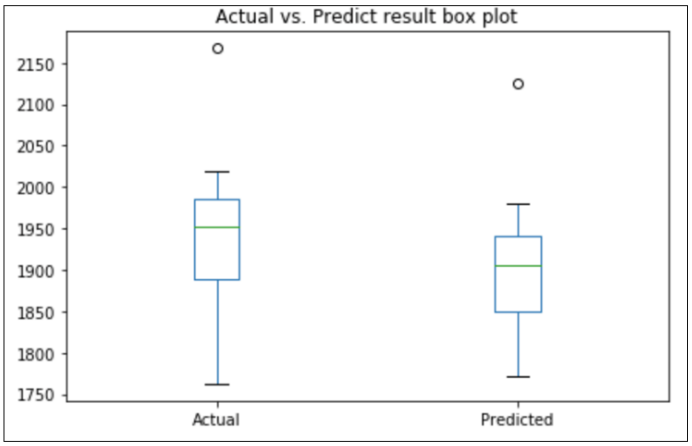
```
Performance Evaluation
Mean Absolute Error: 65.79307481977031
Mean Squared Error: 5586.5223558637435
Root Mean Squared Error: 74.74304219031858
```

    - The performance of ANN:

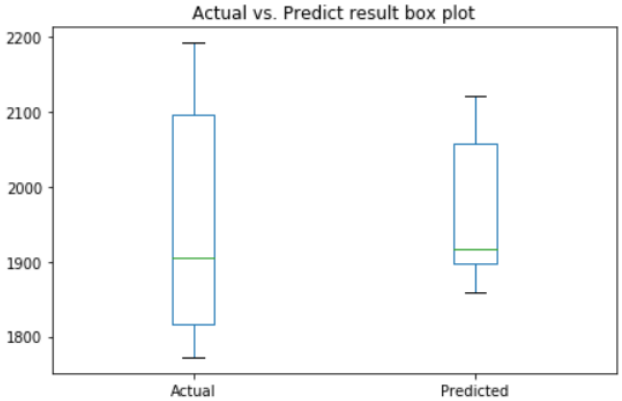
```
Performance Evaluation
Mean Absolute Error: 57.88461100260417
Mean Squared Error: 5276.630399098993
Root Mean Squared Error: 72.64041849479526
```

    - Compare to regression analysis, the MAE, MSE, and RMSE decrease a bit, which means that the model performs better than the regression model.

  - v. Box plot Comparison
    - Regression Analysis box plot



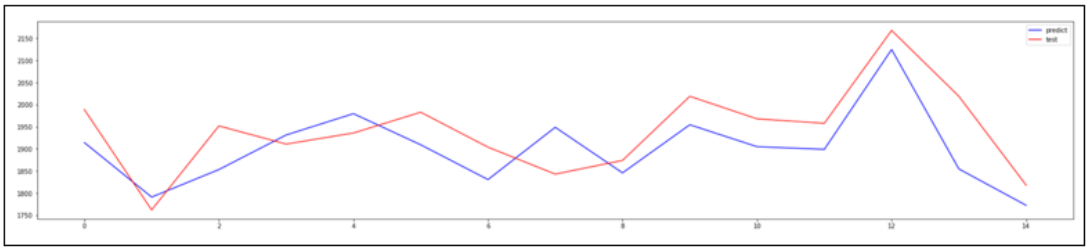
    - ANN box plot



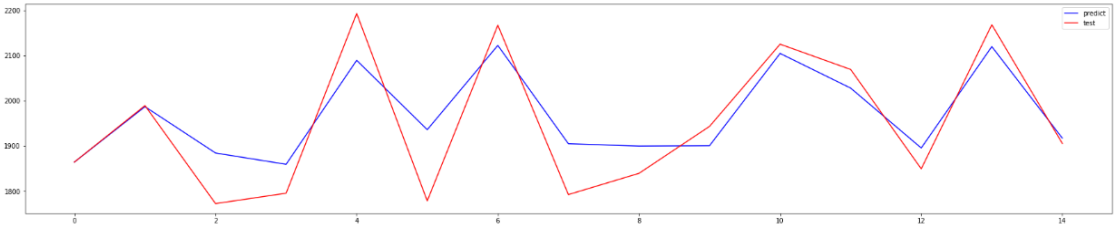
From the box plots, we can see that both models perform well on predicting commercial electricity consumption in Minnesota. However, the ANN model is unable to predict the variance in data and the regression model has larger bias than ANN.

vi. Line plot Comparison

- Regression Analysis Line plot



- ANN Line plot



From the line plots, they show that both models do not fit the data well.

By comparing the performance of the two models, we can conclude that both regression model and ANN are able to predict the commercial industrial consumption in Minnesota. However, it needs more time to find our more related features and tuning the hyper parameters.