Forecasting Electric Vehicle Market Dynamics

Team:
Afrin Unnisa Syed
Somesh Oza
Kumar Shivam Singh
Vinay Krishna Kumar

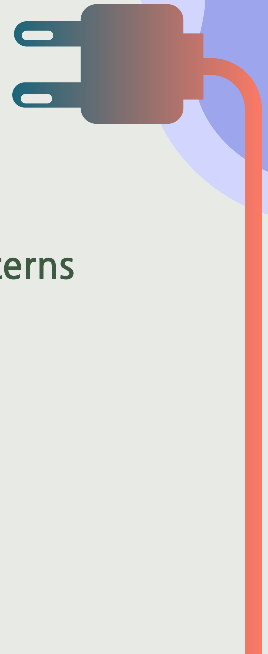
Motivation

- Increasing adoption of Electric Vehicles (EVs) driven by environmental concerns, fuel price volatility, and consumer preferences.

- Understanding EV adoption dynamics is critical for informed decision-making by stakeholders (Government ,manufacturers, consumers).

- Leveraging statistical and machine learning models to uncover patterns and correlations influencing EV adoption.

- The project aims to bridge knowledge gaps and contribute to sustainable transportation planning.



Introduction

- The automotive industry is undergoing a significant transformation with the increasing adoption of electric vehicles (EVs).

- Our project aims to understand and predict these market dynamics through sophisticated statistical analysis.

- Through time series analysis, including ARIMA and SARIMA models, and regression techniques (Ridge, Lasso, and ElasticNet) and Bayesian probability analysis, we've uncovered significant correlations and patterns.

Data

EV Registrations

Year_Month	Fuel_Category	County	Count
2020/07	Electric	ALLEGANY	29
2020/07	Electric	ANNE ARUNDEL	1,587
2020/07	Electric	BALTIMORE	1,508
2020/07	Electric	BALTIMORE CITY	770
2020/07	Electric	CALVERT	140
2020/07	Electric	CAROLINE	8
2020/07	Electric	CARROLL	292
2020/07	Electric	CECIL	82
2020/07	Electric	CHARLES	197
2020/07	Electric	DORCHESTER	24
2020/07	Electric	FREDERICK	710
2020/07	Electric	GARRETT	15
2020/07	Electric	HARFORD	321
2020/07	Electric	HOWARD	1,934

Fuel Prices

Report Date	Gasoline	E85	CNG	LNG	Propane	Diesel	B20	B99/B100
4/10/2000	$1.52	$1.80	$0.89		$1.62	$1.29		
10/9/2000	$1.54	$1.90	$1.02		$1.76	$1.46		
6/4/2001	$1.68	$1.85	$1.30		$1.72	$1.37		
10/22/2001	$1.27	$1.60	$1.19		$1.62	$1.19	$1.35	
2/11/2002	$1.11	$1.54	$1.09		$1.62	$1.04	$1.18	
4/15/2002	$1.40	$1.80	$1.07		$1.95	$1.19	$1.28	
7/22/2002	$1.41	$1.81	$1.20		$1.55	$1.18	$1.39	
10/28/2002	$1.44	$1.71	$1.17		$1.66	$1.35	$1.47	
2/3/2003	$1.61	$1.86	$1.20		$2.09	$1.50	$1.57	
12/1/2003	$1.48	$1.70	$1.35		$2.21	$1.34	$1.60	
3/3/2004	$1.74	$1.84	$1.40		$2.48	$1.47	$1.61	
6/14/2004	$1.99	$2.28	$1.40		$2.13	$1.55	$1.89	
11/15/2004	$1.97	$2.30	$1.56		$2.91	$1.93	$2.05	
3/21/2005	$2.11	$2.29	$1.56		$2.65	$2.03	$2.11	
9/1/2005	$2.77	$3.21	$2.12		$3.50	$2.54	$2.67	$3.30
1/1/2006	$2.23	$2.65	$1.99		$2.71	$2.32	$2.42	$3.14

Exploratory Data Analysis

EV Registrations

Summary Table:

	Column	Data Type	Missing Values	Unique Values	Most Frequent
0	Year_Month	object	0	51	2024/09
1	Fuel_Category	object	0	3	Electric
2	County	object	16	568	ALLEGANY
3	Count	object	0	1254	1

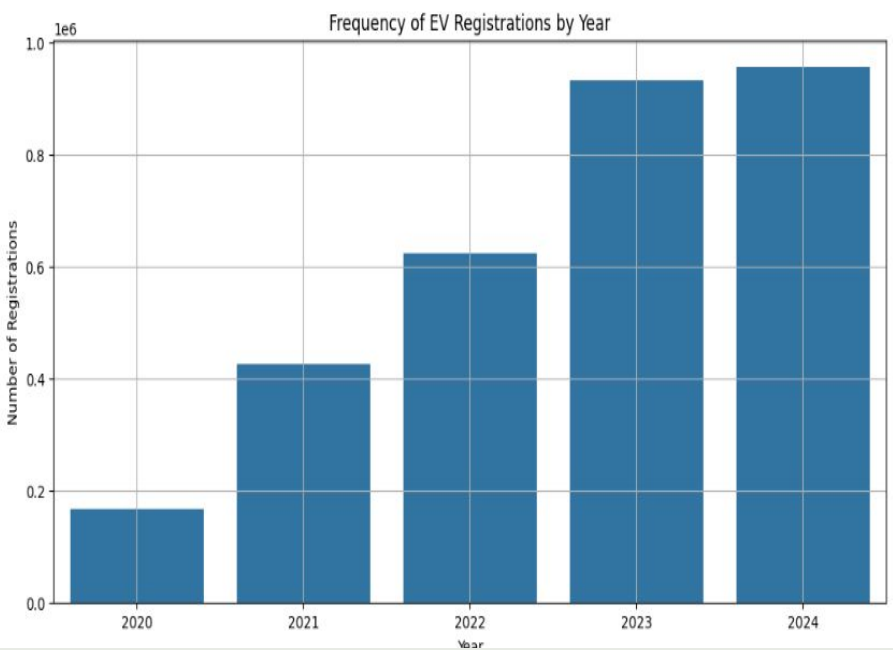
Fuel Prices

Summary Table:

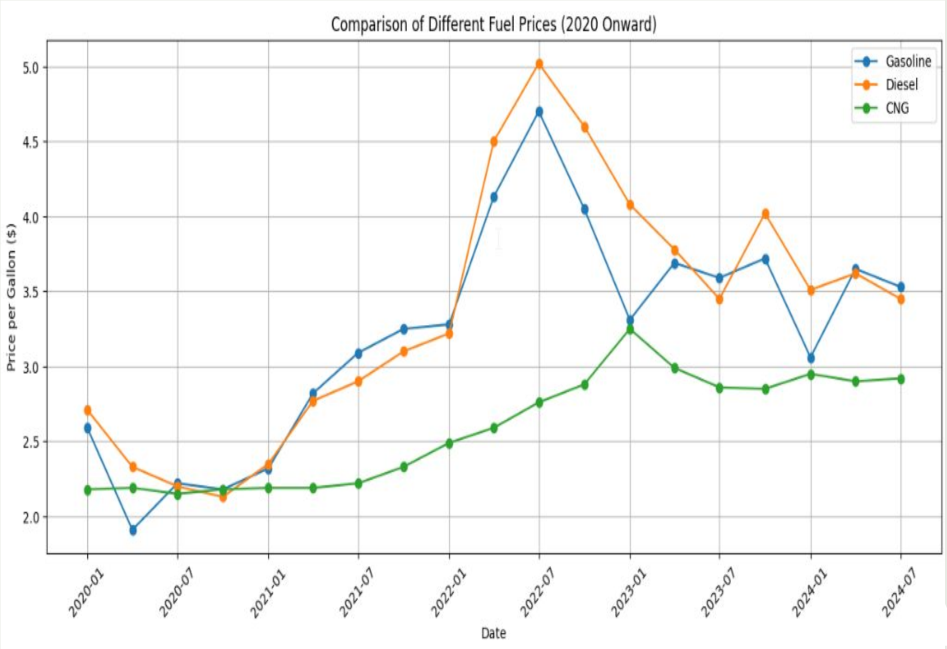
	Column	Data Type	Missing Values	Unique Values	Most Frequent
0	Report Date	object	0	88	1/1/06
1	Gasoline	object	0	74	$2.22
2	E85	object	0	77	$2.65
3	CNG	object	0	56	$2.09
4	LNG	object	55	28	$2.40
5	Propane	object	0	68	$1.62
6	Diesel	object	0	73	$2.71
7	B20	object	3	71	$2.11
8	B99/B100	object	14	63	$3.65

Exploratory Data Analysis

EV Registrations



Fuel Prices



Regression: Data Preprocessing

- **Monthly Aggregation**: Group the EV registration data by 'Year_Month' and sum up the 'Count' values for each group. This aggregates the data on a monthly basis, providing a consolidated view of monthly EV registration.
- **Standardize Date Format**: Convert the 'Report Date' in the fuel prices dataset and 'Year_Month' in the aggregated EV registration data into pandas DateTime format. This ensures that the dates are in a uniform format, facilitating easier manipulation and merging based on date fields.
- **Merge Based on Dates**: Merge the monthly EV registration data with the fuel prices data using merge, which matches each record from the EV data to the last available record from the fuel data before or on the matching date. This is crucial for ensuring that the fuel prices are appropriately aligned with the EV registration dates, considering the closest available report date.

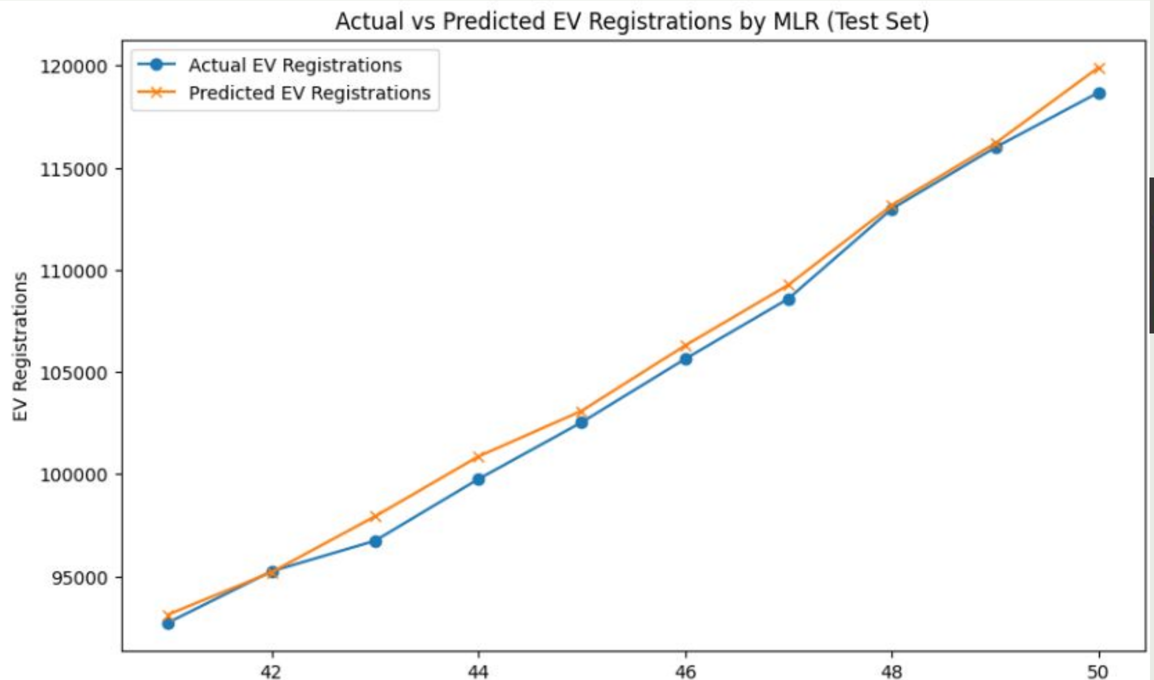
Regression: Feature Engineering

- **Lagged Features**: **1-Month Lag**: Create a new feature `EV_Count_Lag1` by shifting the 'Count' column down by one period (one month). This introduces a lagged feature representing the number of EV registrations from the previous month.
- **Rolling Averages**: **3-Month Rolling Average**: Calculate a rolling average of the 'Count' over a window of three months and store it in `EV_Count_Rolling3`. This helps in smoothing out short-term fluctuations and reveals underlying trends in the data.
- **Input Features**: The final set of features used for modeling includes 'Gasoline', 'Diesel', 'CNG' prices, the 1-month lagged EV registrations (`EV_Count_Lag1`), and the 3-month rolling average of EV registrations (`EV_Count_Rolling3`).
- **Target Variable**: The target variable is the 'Count' of EV registrations.

Regression

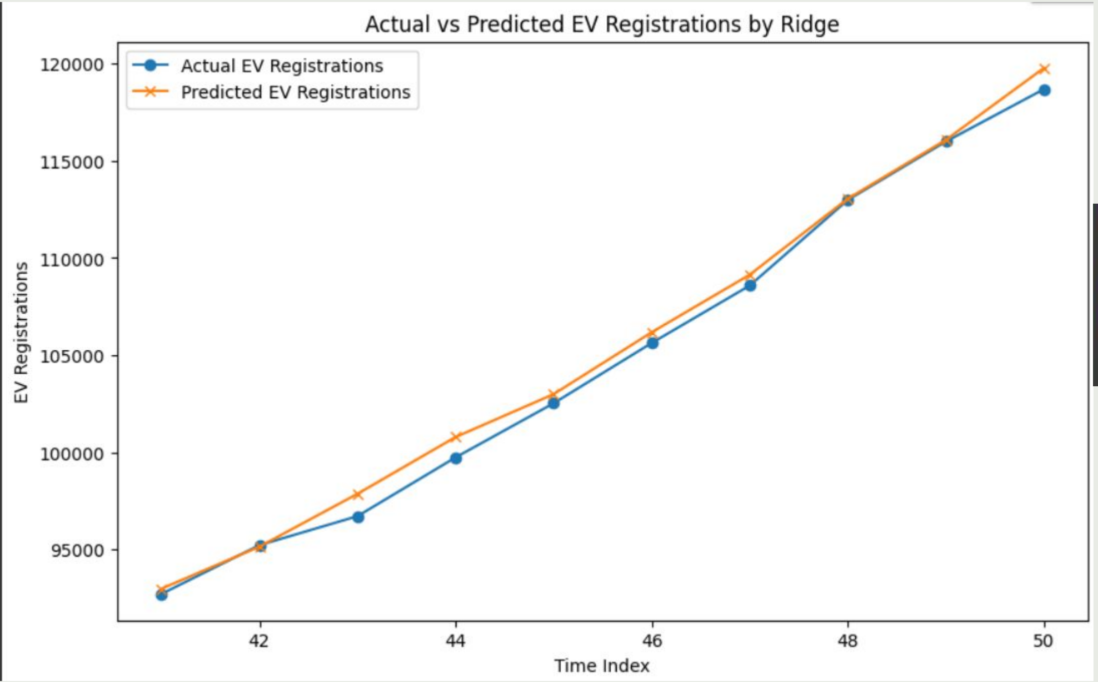
- **Multiple Linear Regression (MLR)** is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- **Lasso** (Least Absolute Shrinkage and Selection Operator) regression is another type of linear regression that uses shrinkage. Lasso regression performs L1 regularization, which **adds a penalty equal to the absolute value of the magnitude of coefficients.**
- **Ridge regression** is a technique used to analyze data that suffer from multicollinearity. When independent variables are highly correlated, a small change in one variable might be associated with a high degree of change in another, thus resulting in calculation difficulties.**It adds a penalty equal to the square of the magnitude of coefficients to the loss function (sum of squared residuals).**
- **Elastic Net** is a hybrid of Ridge and Lasso regression. It integrates the penalties of both models, combining L1 and L2 regularization. It is useful when there are multiple features correlated with one another.

Results : Multiple Linear Regression



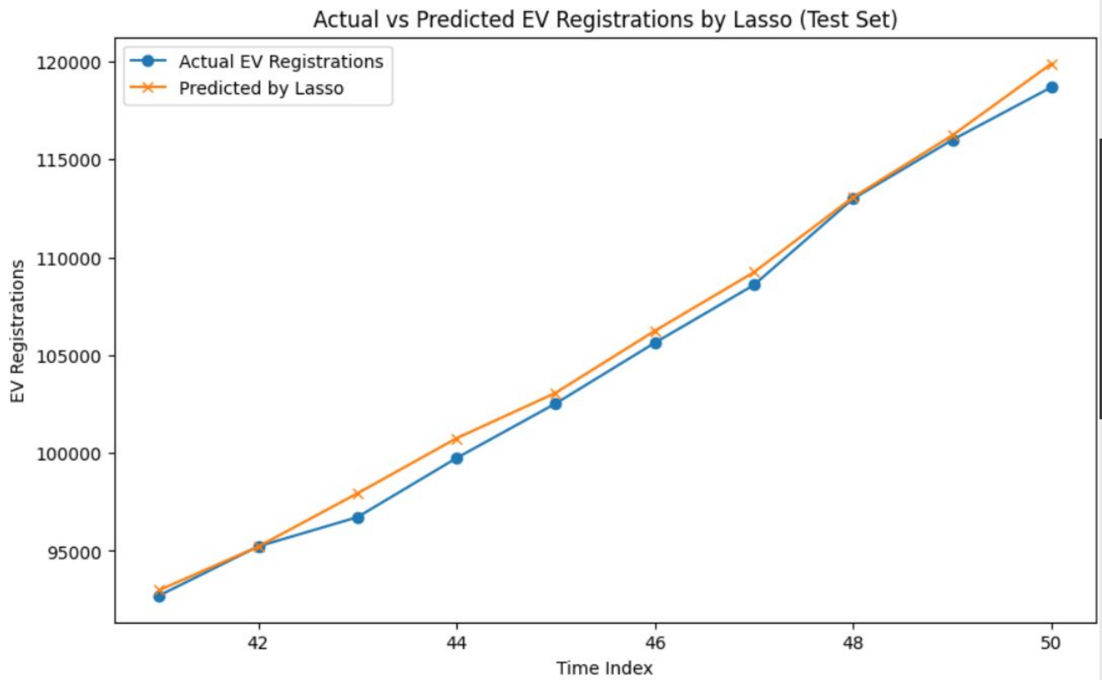

```
Multiple Linear Regression Results:
R-squared (Train): 0.9997
R-squared (Test): 0.9924
Test RMSE: 747.91
```

Results : Ridge Regression



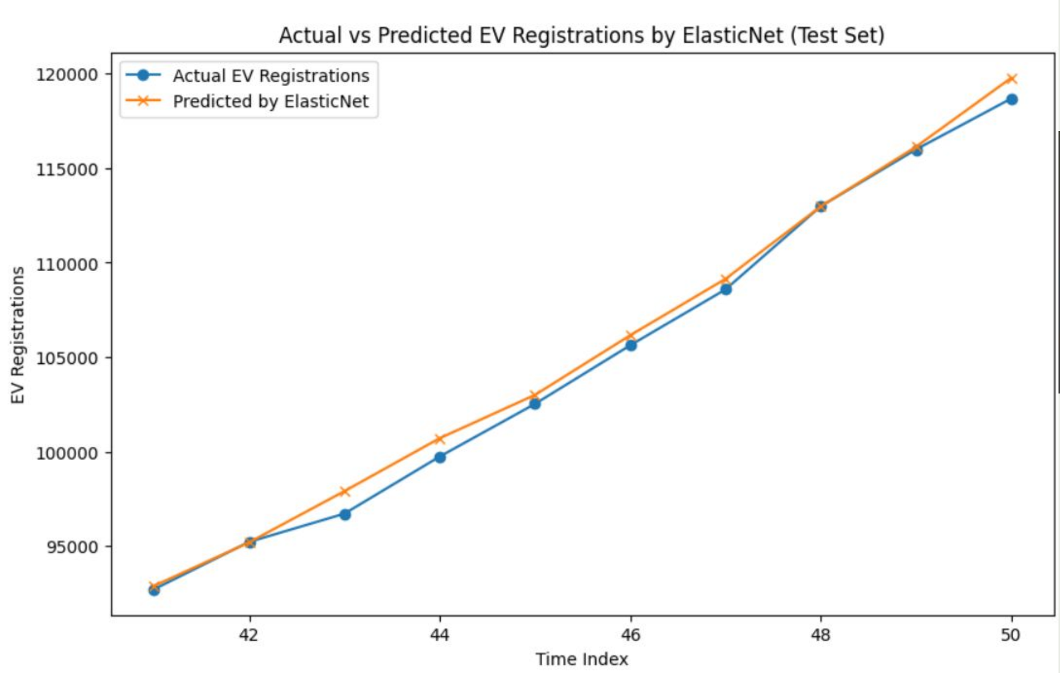

```
Ridge Regression with Cross-Validation Results:
Optimal alpha: 21.54434690031882
R-squared (Train): 0.9997
R-squared (Test): 0.9938
Test RMSE: 672.64
```

Results : Lasso Regression




```
Lasso Regression Results:
Optimal alpha (Lasso): 1.0
R-squared (Train - Lasso): 0.9997
R-squared (Test - Lasso): 0.9930
Test RMSE (Lasso): 718.17
```

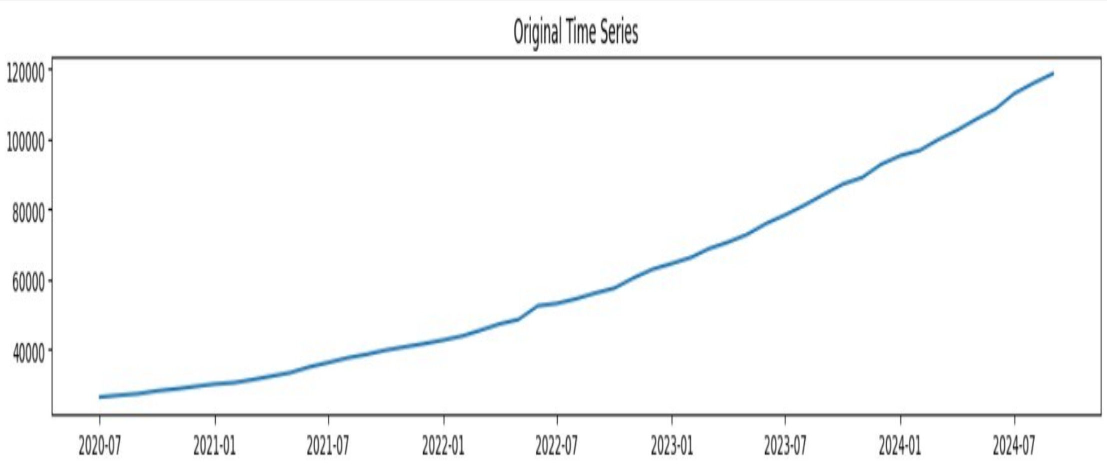
Results : ElasticNet Regression




```
ElasticNet Regression Results:
Optimal alpha (ElasticNet): 1.0
Optimal l1 ratio (ElasticNet): 0.5
R-squared (Train - ElasticNet): 0.9997
R-squared (Test - ElasticNet): 0.9940
Test RMSE (ElasticNet): 664.95
```

ARIMA(1, 1, 1) Model

Goal: Analyze the trend, seasonality, and forecast future electric vehicle (EV) registrations.



Why It Matters:

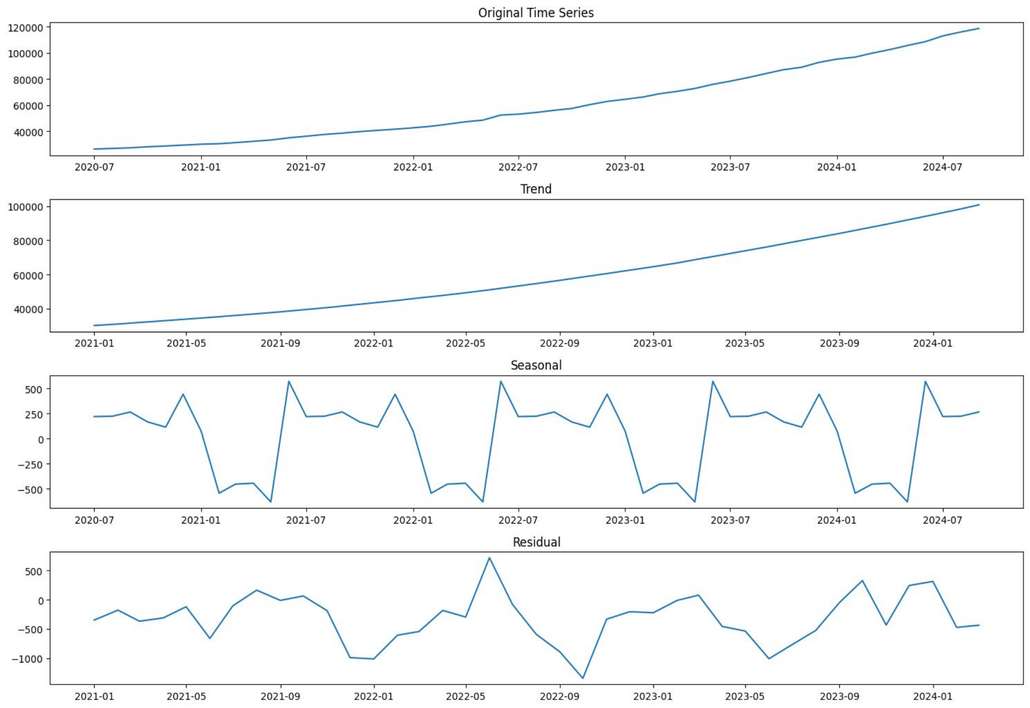
- EV registrations show strong growth, and accurate forecasts are critical for planning.
- Understanding trends and patterns helps businesses, policymakers, and infrastructure planners.

Preprocessing Steps:

- Aggregated total registrations by month.
- Converted the `Year_Month` column to a **Datetime Index**.
- Cleaned the data to ensure numerical consistency.

ARIMA (1, 1, 1) Model

Time Series Decomposition of EV Registrations: Trend, Seasonality, and Residuals



Trend: Long-term growth pattern.
Seasonality: Repeating cycles or patterns.
Residuals: Irregular/random fluctuations after removing trend and seasonality.

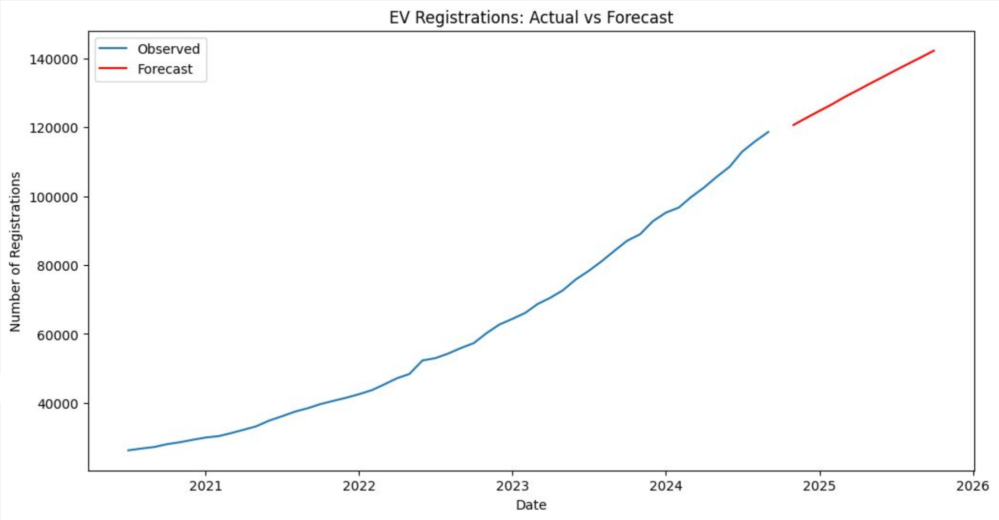
ARIMA (1, 1, 1) Model

Test Used: Augmented Dickey-Fuller (ADF) test.

```
Augmented Dickey-Fuller Test:
ADF Statistic: 9.109480264705978
p-value: 1.0
```

P> 0.05: Data is **not stationary**

Forecasting Results:

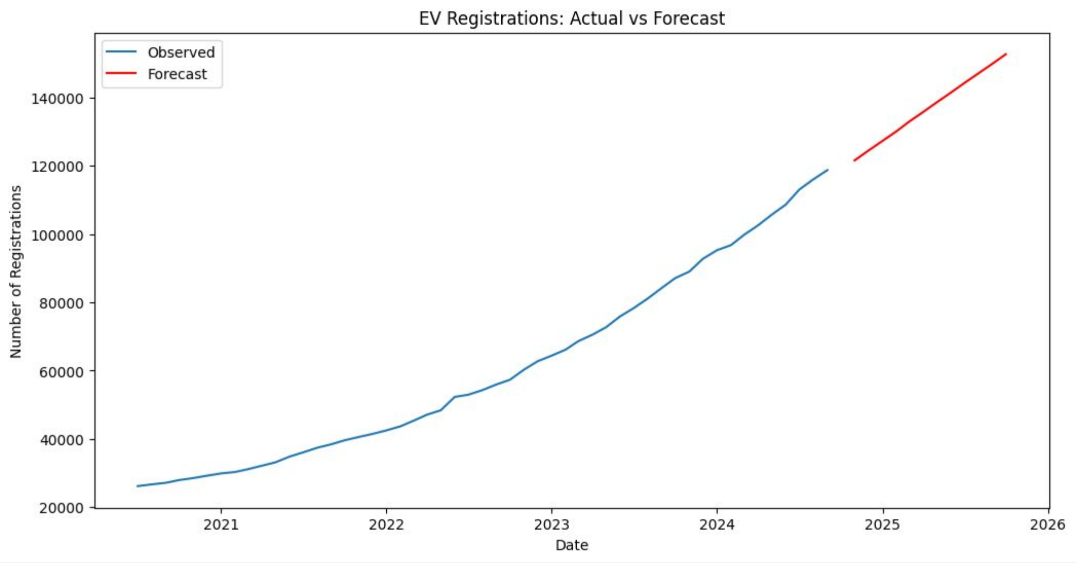


Parameters:

- p=1: One lag of past data (AR term).
- d=1: First differencing to make the data stationary.
- q=1: One lagged error term (MA term).

ARIMA Model Using Auto-Arima

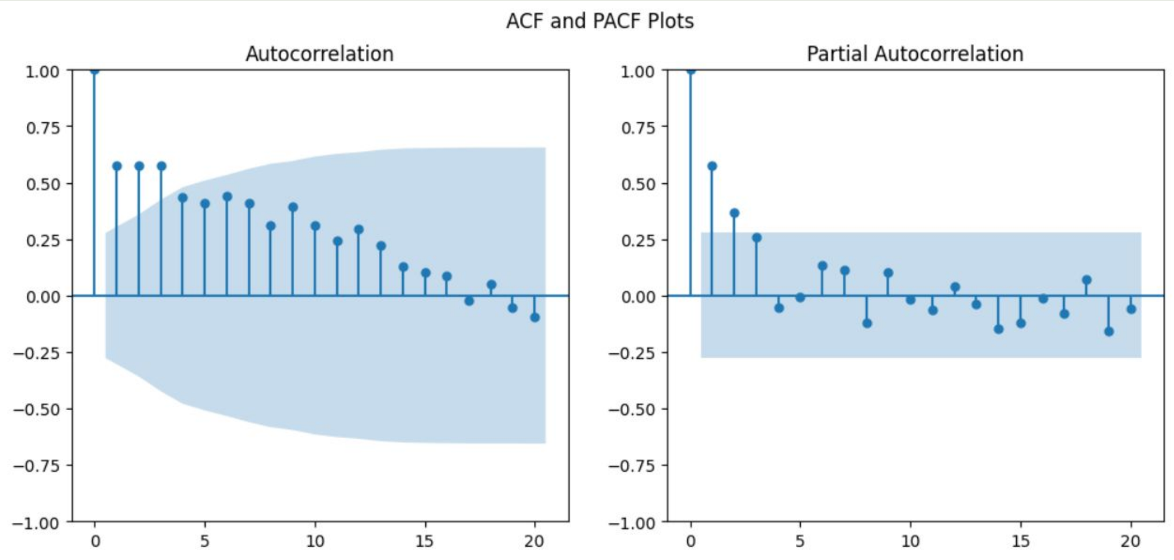
```
Best ARIMA Order: (0, 2, 1)
```



Parameters:

- p=0: One lag of past data (AR term).
- d=2: First differencing to make the data stationary.
- q=1: One lagged error term (MA term).

ARIMA Model Using Auto-Arima



X-axis:

- Represents the **lag** values, which are the number of time steps back (e.g., 1 month ago, 2 months ago, etc.).

Y-axis:

- Represents the **correlation coefficient**

PACF bars are within the confidence interval (shaded region) beyond lag 0.

ARIMA Model Comparison

```
ARIMA Model Summary:
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Count   No. Observations:                   51
Model:                 ARIMA(0, 2, 1)   Log Likelihood                -400.171
Date:                Tue, 17 Dec 2024   AIC                            804.343
Time:                        21:14:56   BIC                            808.126
Sample:                    07-01-2020   HQIC                           805.778
                         - 09-01-2024
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.2369      0.045     -5.229      0.000      -0.326      -0.148
sigma2      6.445e+05   9.14e+04      7.054      0.000    4.65e+05    8.24e+05
===================================================================================
Ljung-Box (L1) (Q):                   7.92   Jarque-Bera (JB):                10.02
Prob(Q):                              0.00   Prob(JB):                         0.01
Heteroskedasticity (H):               2.69   Skew:                            -0.07
Prob(H) (two-sided):                  0.06   Kurtosis:                         5.21
===================================================================================
```

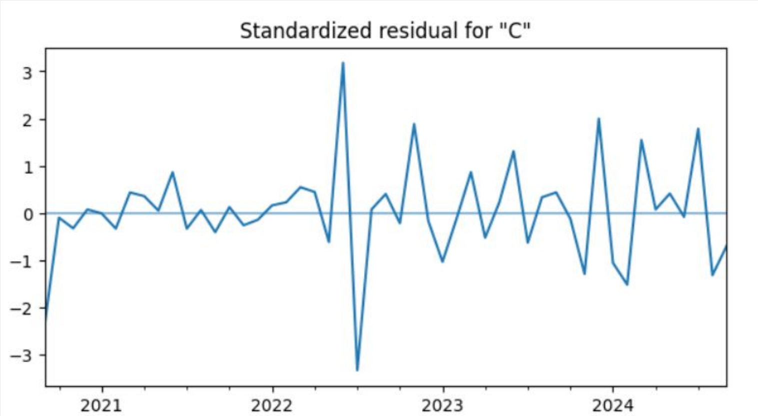
```
ARIMA Model Summary:
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Count   No. Observations:                   51
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -417.324
Date:                Sun, 15 Dec 2024   AIC                            840.647
Time:                        17:09:48   BIC                            846.383
Sample:                    07-01-2020   HQIC                           842.831
                         - 09-01-2024
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9942      0.015     66.090      0.000       0.965       1.024
ma.L1         -0.9545      0.061    -15.738      0.000      -1.073      -0.836
sigma2      9.174e+05    8.7e-09   1.05e+14      0.000    9.17e+05    9.17e+05
===================================================================================
Ljung-Box (L1) (Q):                  11.07   Jarque-Bera (JB):                 0.75
Prob(Q):                              0.00   Prob(JB):                         0.69
Heteroskedasticity (H):               3.60   Skew:                             0.30
Prob(H) (two-sided):                  0.01   Kurtosis:                         2.93
===================================================================================
```

ARIMA(0,2,1) has significantly lower **AIC** (804.343) and **BIC** (808.126) compared to ARIMA(1,1,1).

This suggests ARIMA(0,2,1) is a better model in terms of goodness-of-fit.

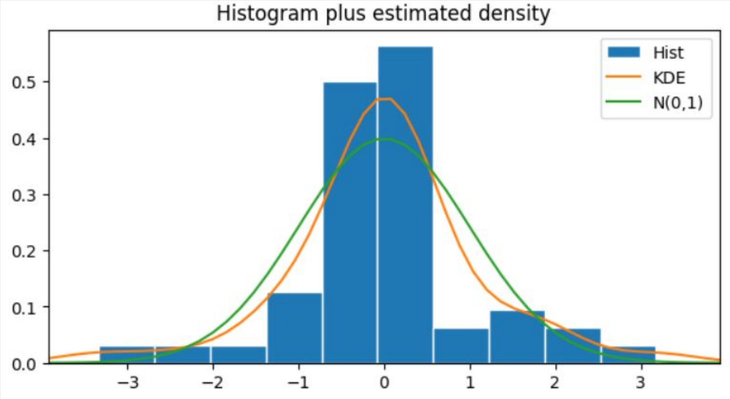
SARIMA(0, 2, 1, 12)

Standardized Residuals



- The residuals hover around **zero** without a clear trend.
- Some spikes (outliers) are observed, especially in 2022 and 2023, indicating minor irregularities.
- There is **no strong autocorrelation**

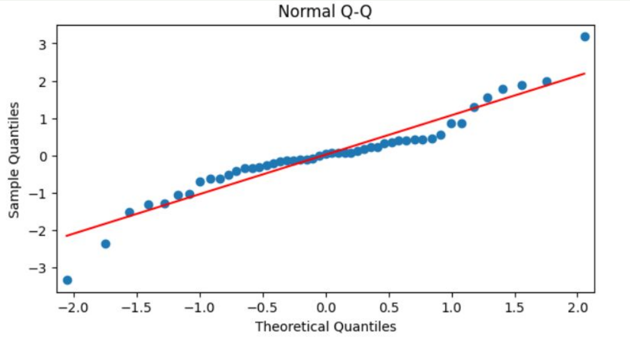
Histogram with KDE



- The histogram is roughly **centered around zero.**
- The KDE curve aligns closely with the normal distribution, indicating that the residuals are approximately **normally distributed**.

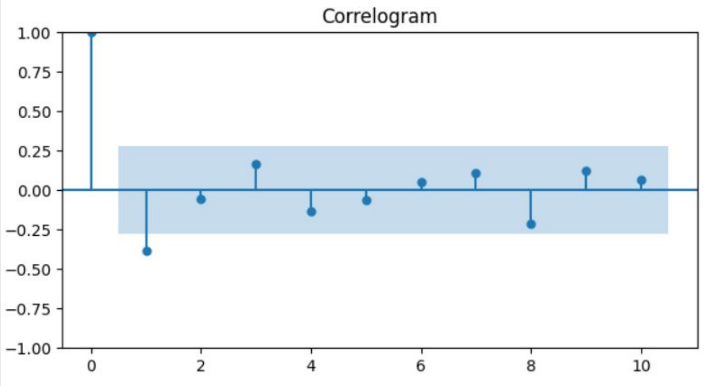
SARIMA(0, 2, 1, 12)

Normal Q-Q (Quantile-Quantile) Plot



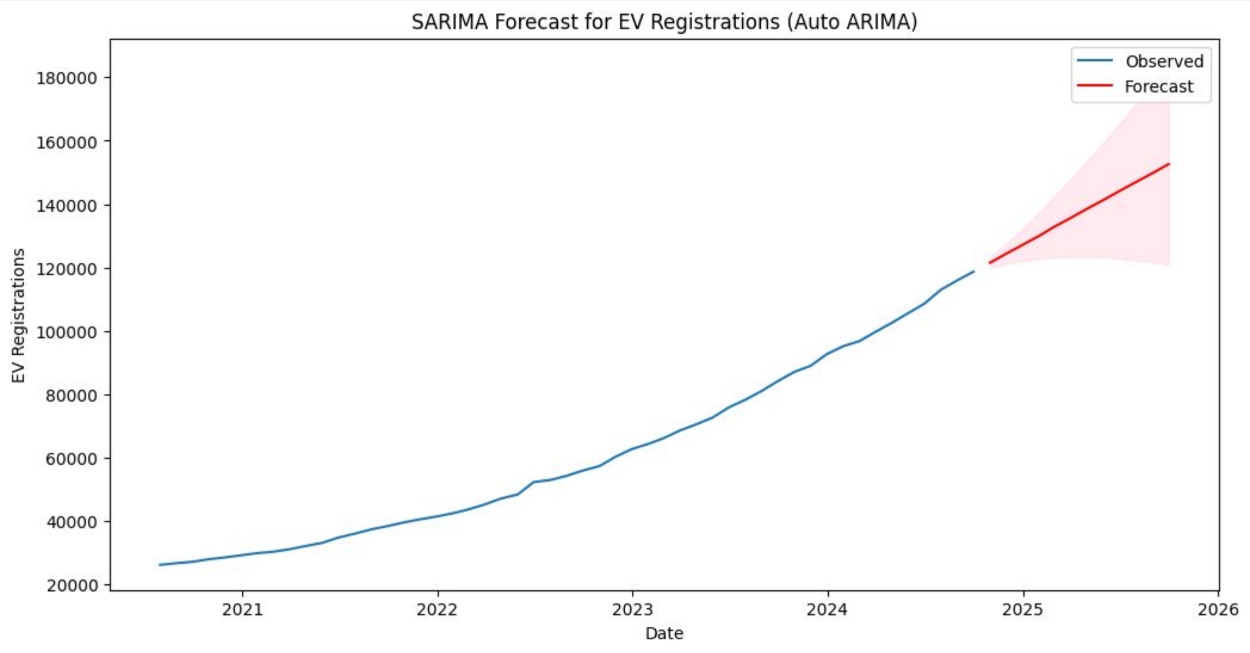
- Residual points (blue dots) lie **close to the red diagonal line**, which represents the ideal normal distribution.

Correlogram (ACF of Residuals)



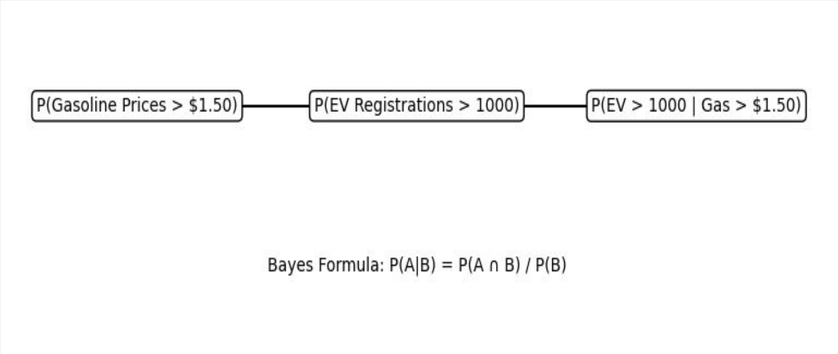
- Most autocorrelations are **within the shaded region**, indicating **no significant autocorrelation**.
- This suggests that the residuals are random and do not contain any further information that the model missed.

SARIMA (0, 2, 1, 12) Forecast



Bayesian Analysis of EV Registration and Gasoline Prices

- What happens to electric vehicle (EV) registrations when gasoline prices rise?
- How can we calculate the probability of EV registrations exceeding 1,000 given that gasoline prices are above $1.50?
- How can Bayes' Theorem help us update this probability as new information comes in?




```
Prior Probability P(A): 0.4
Prior Probability P(B): 0.6
Conditional Probability P(B|A): 1.0
Posterior Probability P(A|B): 0.6666666666666667
```

Challenges

- Minimal dataset overlap limited model generalization and diversity
- Finding optimal p, d, and q was difficult, so auto ARIMA was used.
- Overconfidence in Conditional Probability
- Due to the limited dataset, there is a possibility of overfitting.

Conclusion

- The MLR model demonstrates strong predictive performance, with high R^2 and a low RMSE, making it suitable for predicting the target variable with confidence.
- The SARIMA and ARIMA models performed similarly, with minimal difference in fit and predictive accuracy.
- Bayes' Theorem helped estimate the likelihood of increased EV registrations with higher gasoline prices, showing a 66.7% chance but not a certainty.

THANKS!