# Pipeline Incident Analysis (2010 - 2017)
## Analysis and Forecast Modeling

Evan Kiolbassa

## Introduction

One of the greatest challenges pipeline operators face today is public opinion. One notable event in the recent past is the opposition of the Keystone XL pipeline construction in South Dakota. To mitigate public relations disasters, ensure employee and surrounding communities safety, and minimize incident related costs, it is vital to implement robust asset reliability systems. These include leak detection systems, predictive maintenance, sophisticated SCADA systems, and computational modeling of corrosion. The purpose of

this project is to analyze incident cause, frequency, and associated losses. The development of a SARIMAX model for monthly incident forecasting is also discussed.

# Evaluation of Missingness and Data Cleaning

The data that is analyzed in this report is sourced from the Kaggle dataset "Oil Pipeline Accidents, 2010 to Present." For a more detailed description of the data, please refer to the appendix section. Figure 1 is a missingness matrix from the msno package. There is a large amount of missingness in the dataset, but after careful evaluation it is all a result of the value being zero or invalid. For integer values, zero was imputed. For datetime variables such as "Restart/Shutdown Time," the accident date was imputed leading to a time delta of zero.

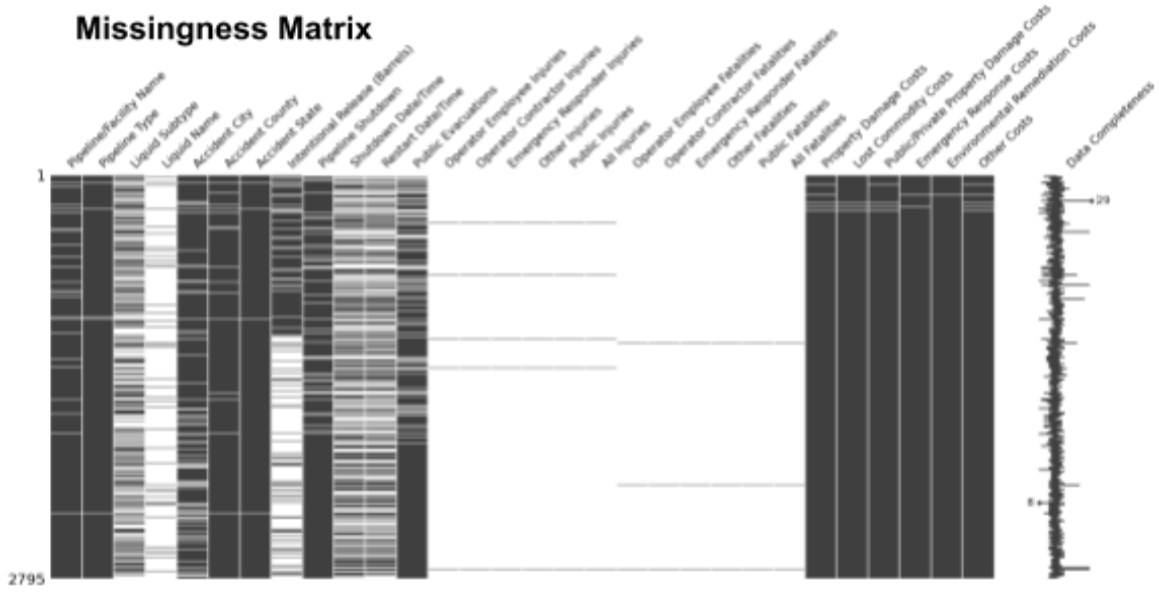


Figure 1: Missingness Matrix

# Exploratory Data Analysis

## Yearly Trend Analysis

Figure 2 shows the trend of yearly incidents from 2010 to 2016. There is polynomial growth from 2011 to 2014 where the incident rate remained positive but decreased significantly to 2015. The incident rate dropped significantly from 2015 to 2016.

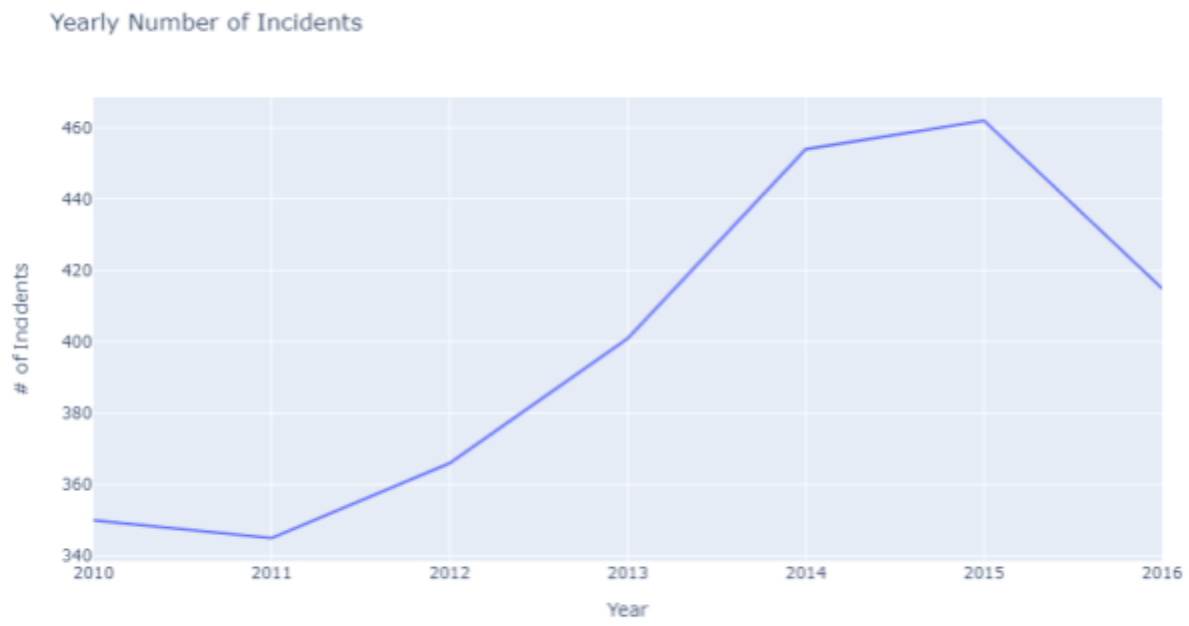


Figure 2: Yearly Incident Trend

Figure 3 shows the trends of yearly incidents by cause category. It is evident that the primary cause of pipeline incidents is material/weld/equipment failure. Corrosion and incorrect operation are also significant contributors. These insights solidify the importance of reliability, process safety management, and strict compliance to standard operating procedures. This would suggest the need for more rigid engineering controls preventing operators from by-passing certain safety/interlock features.

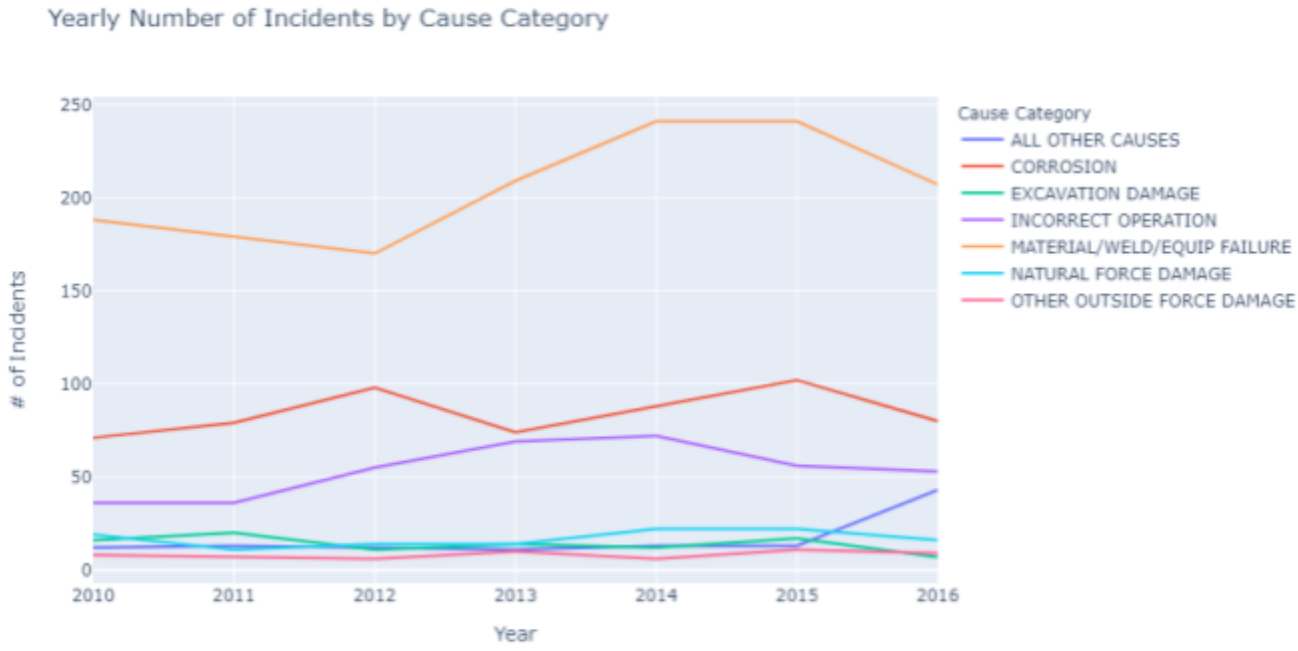


Figure 3: Yearly Incident Trend by Cause Category

While equipment related failure is the most frequent mode of incident cause, Figure 4 shows that internal corrosion is the most frequent cause subcategory while external corrosion is the fourth highest cause. The most common modes of equipment failure are pump-related and non-threaded connection failures (weld-on/torqued flanges). This exposes the vital importance of predictive modeling of corrosion to minimize asset losses. Under the assumption that flanges are properly torqued to proper specifications, this may suggest an element of seasonality to connection and equipment failure in general. This will be explored later in the section on time series analysis/modeling.

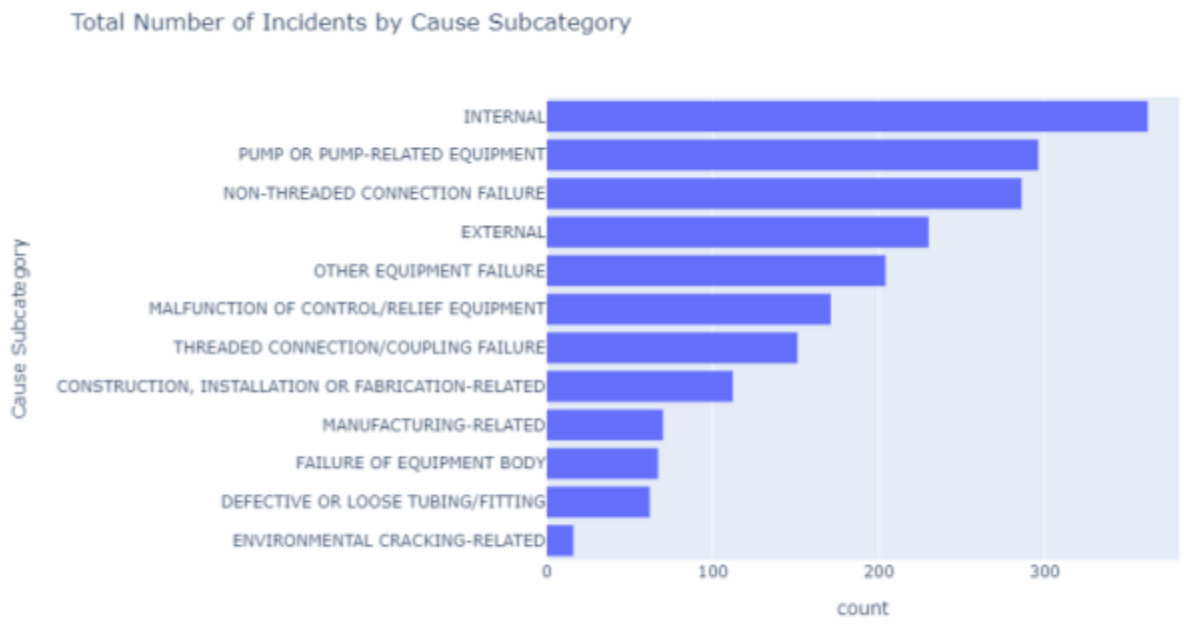


Figure 4: Bar Chart of Total Number of Incidents Associated with Each Incident Subcategory

## Incident Cost Analysis

The cost distribution in Figure 5 reveals a heavily right-skewed distribution with a high degree of kurtosis. The calculated skew of the cost distribution is 47.06 while the skewness is 2361.06. Most Incidents result in losses less than one-million U.S Dollars, but the highest frequency is under $200,000.

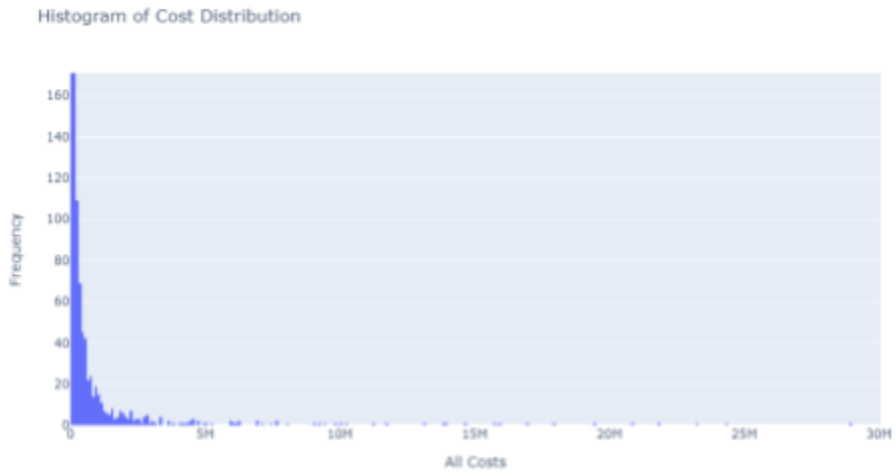


Figure 5: Incident Cost Distribution

Figure 6 shows the ratio of losses associated with equipment/corrosion relative to all losses by year. In the year 2010, 83.4% of total incident losses were attributed to equipment/corrosion cases. On 7-25-2010, the largest net loss within the analyzed time frame occurred in Michigan at an Enbridge Energy asset, resulting in losses of $840,526,118.

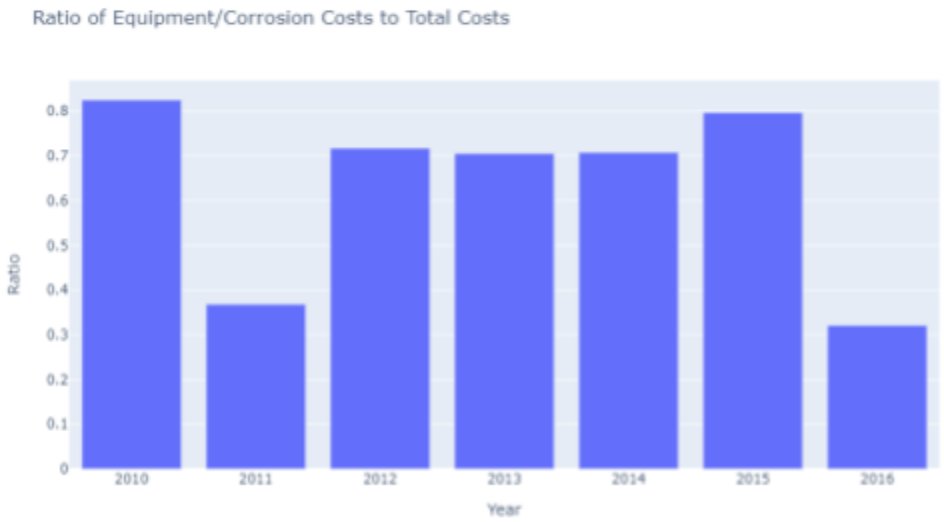


Figure 6: Bar Graph of Equipment /Corrosion Cost Ratios by Year

Figure 7 shows the relationship between incident cost and the net loss of product in barrels after log base 10 transformation. Performing a simple linear regression between the two variables yielded a R-Squared of 0.110, suggesting that there is little predictive power without any additional aggregation or scaling. One method of future exploration could involve method iteration of robust scaler and cross-validate regression performance by adjusted R-Squared. If expanding the scope of this project would involve building a predictive regression model, the change in crude prices over time would also need to be accounted for.


Figure 7: Scatter Plot of Cost per Net Loss in Barrels

Figure 8 shows the distribution of net product loss in barrels. The distribution is heavily right skewed and has severe kurtosis. A majority of incidents (2,496) result in net losses beneath 25 barrels, signifying that most incidents are minor.

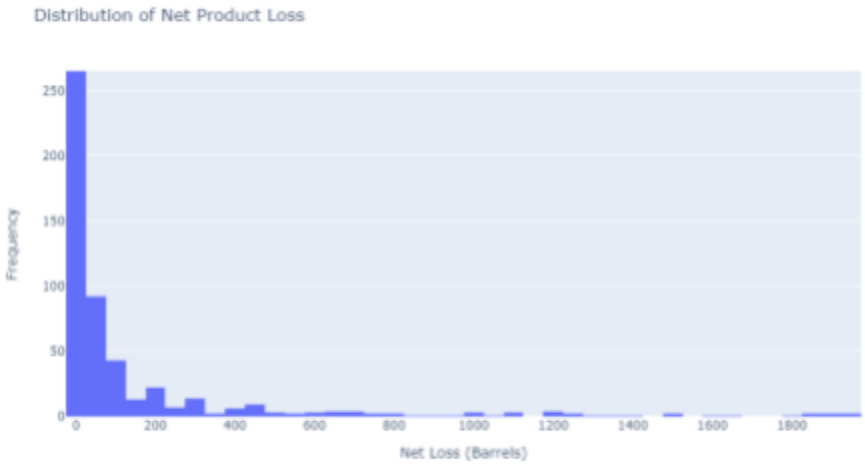


Figure 8: Net Product Loss Distribution

Figure 9 is a log scale scatter plot showing the relationship of price per hour of downtime resulting from incidents. Performing a simple linear regression yielded a R-Squared value of 0.149, but there is a clear relationship between the two variables. Further aggregation by cause category would likely result in more predictive power from a descriptive MLR model.

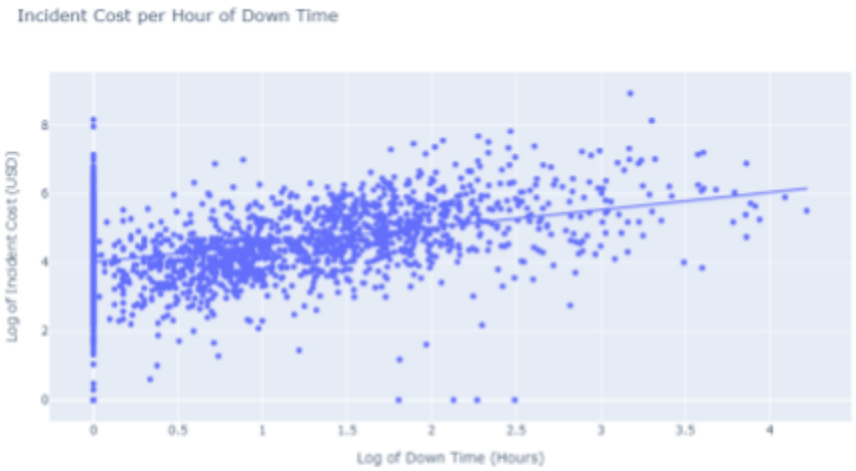


Figure 9: Log Scatter Plot of Price per Hour of Downtime

The distribution of incident downtime, like other numeric variables explored previously, is right skewed with significant kurtosis as visualized in Figure 10. Most incidents result in less than ten hours of downtime. However, there are a significant amount of observations up to 130 hours of downtime.

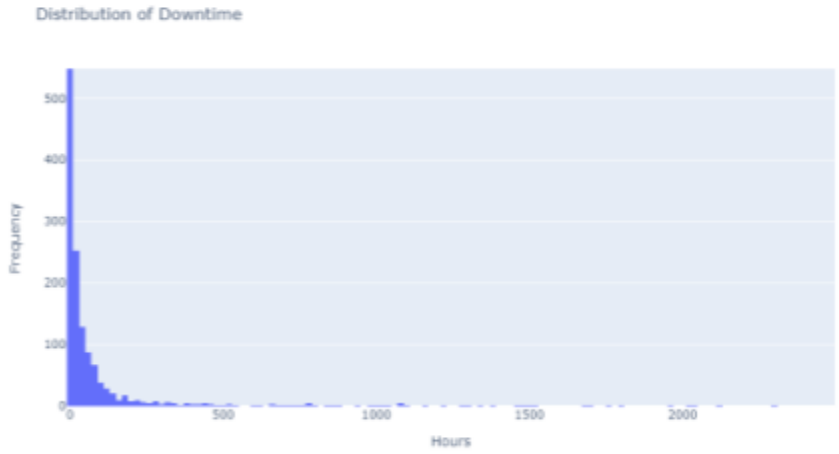


Figure 10: Distribution of Incident Downtime in Hours

## Regional Analysis

Figure 11 visualizes the distribution of incidents by region. Intuitively the Gulf Coast region is a high contributor to incident occurrences due to the density of pipeline assets and refining demand, particularly in Texas and Louisiana.

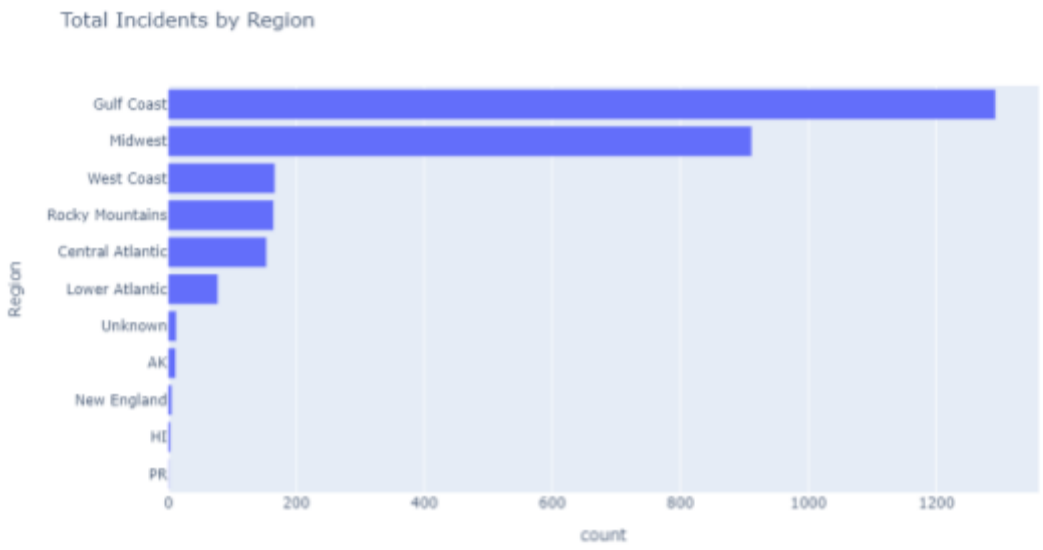


Figure 11: Incident Distribution by Region

Figure 12 is a visualization of the distribution of natural gas pipelines sourced from the U.S Energy Information Administration [1]. The states in the Midwest also have a very high density pipeline infrastructure meaning that it will naturally have a higher incident rate.

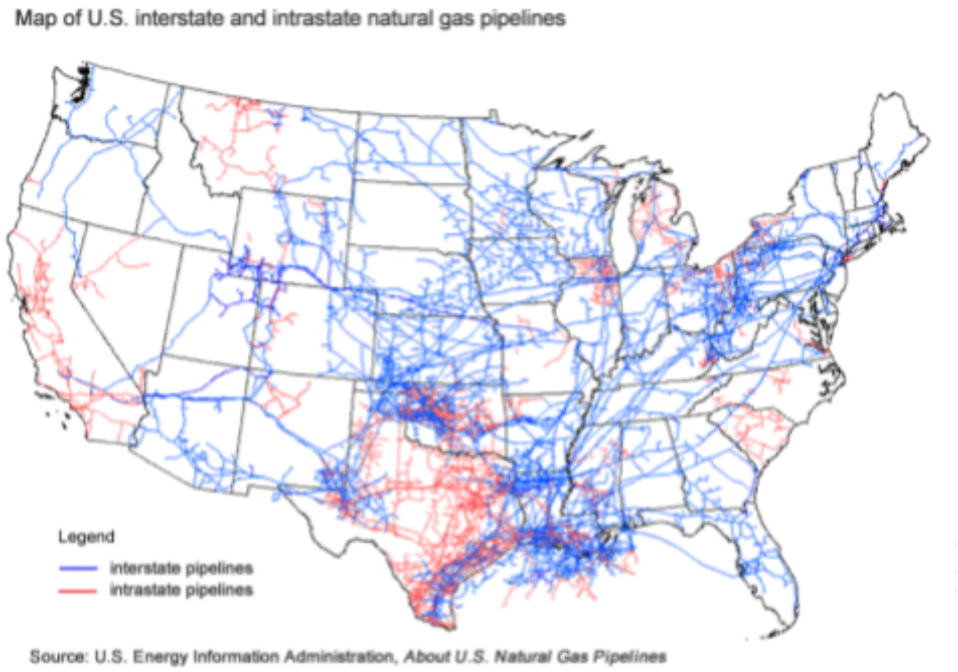


Figure 12: Natural Gas Pipeline Distribution

When adjusting the regional cost to the sum of incident occurrence, Alaska has the highest loss relative to incident rate followed by the Rocky Mountain region.

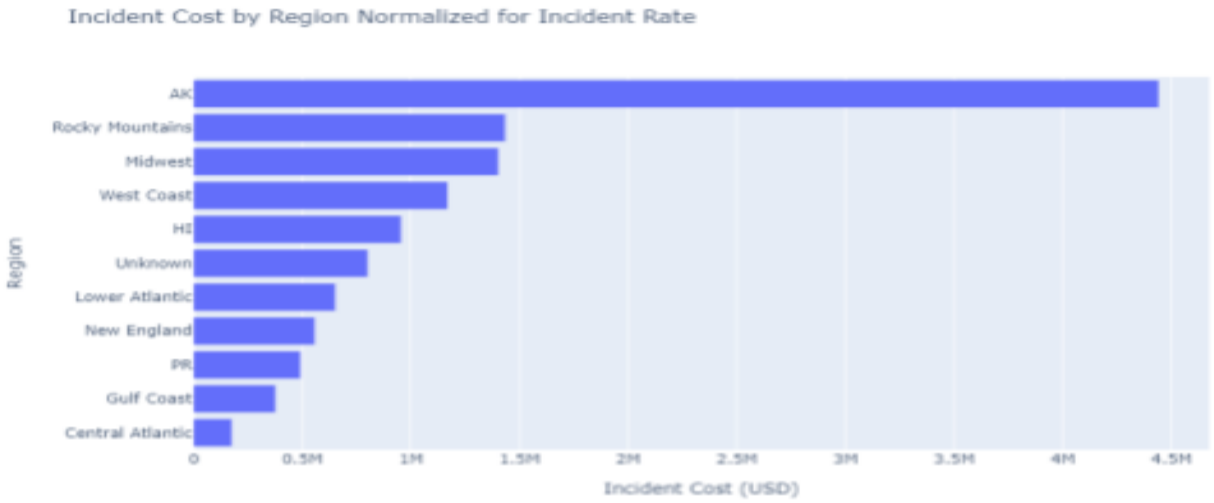


Figure 13: Regional Incident Cost Adjusted for Incident Sum

## Operator Analysis

Figure 14 shows the top twenty operators with the highest incident rate. There are a lot of major oil companies on the list with very high volume assets, however there are a few operators that have lower volume throughput with major losses.

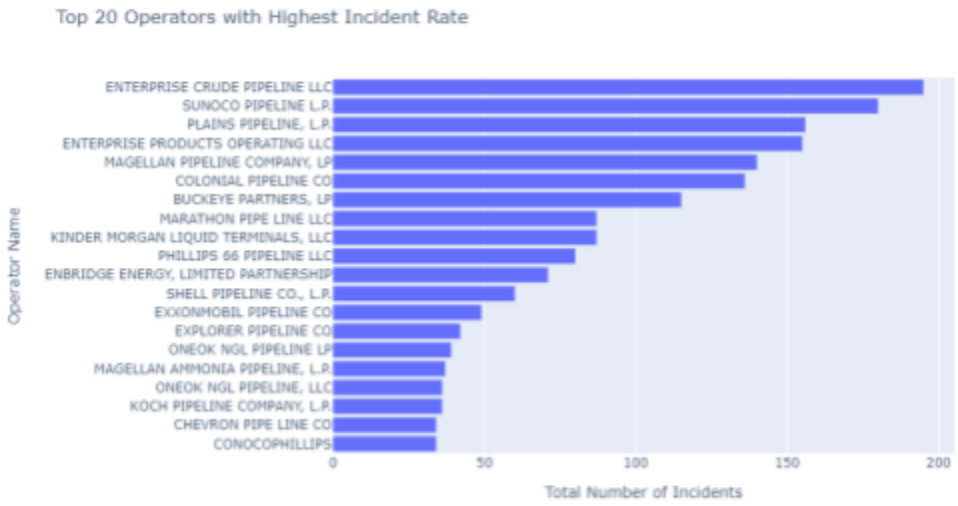


Figure 14: Top 20 Operators with Highest Incident Rate

According to Figure 15, Mobil Pipe Line Company has the highest loss with incident adjustment. Enbridge Energy Limited Partnership has the second highest adjusted loss. A majority of the operators on the adjusted loss list are smaller capacity operators, but there are still some major Fortune 500 companies such as ExxonMobil and Chevron. Operators such as ExxonMobil, Enterprise, Marathon, Koch Industries, and Shell have been expanding their data science capabilities significantly in recent years. Unfortunately, the dataset being explored in this study does not account for return on investment for data science implementation.

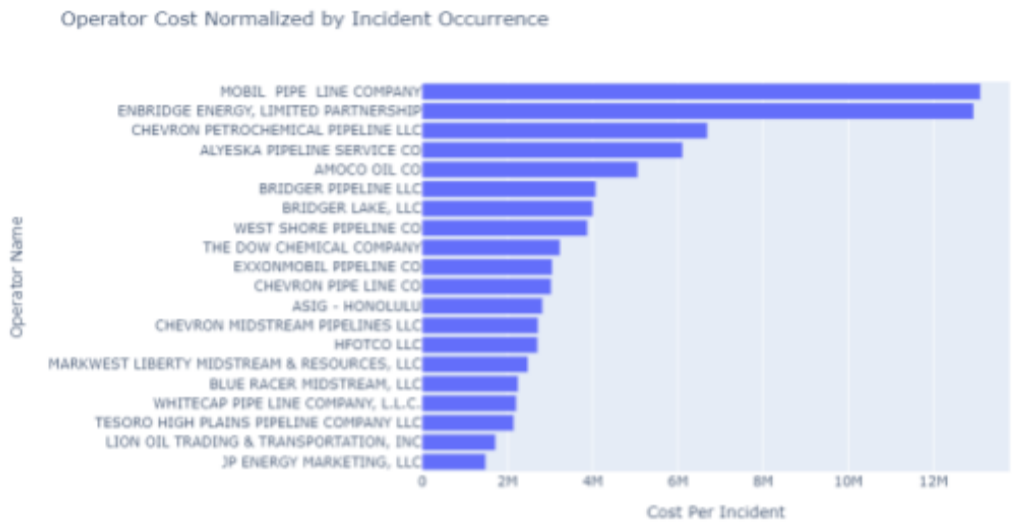


Figure 15: Top 20 Operator Costs Adjusted for Incident Rate

# Time Series Analysis / Modeling

## Data Preparation

For accurate time series modeling, the mean needs to be close to zero, otherwise this suggests the presence of bias. Preliminary models demonstrated that the mean of the time series data was non-zero with a high positive kurtosis. To minimize the effects of skew and kurtosis, a boxcox transformation with robust scaler was applied. The optimal lambda for the transformation is 0.20, thus the following transformation is applied:

`y = (x**lambda - 1) / lambda`

Box-Cox Transformation

## Multiplicative Time Series Decomposition of Equipment Failure Incidents

Multiplicative time series decomposition is a naive method of isolating trend and seasonality from data. This choice of decomposition method was used due to the non-linearity of the data.

y(t) = (Level)(Trend)(Seasonality)(Noise)

Multiplicative Formula

Figure 16 is a visualization of the raw monthly time series data. Based on the raw visualization there does not appear to be any apparent trend component at the monthly frequency.

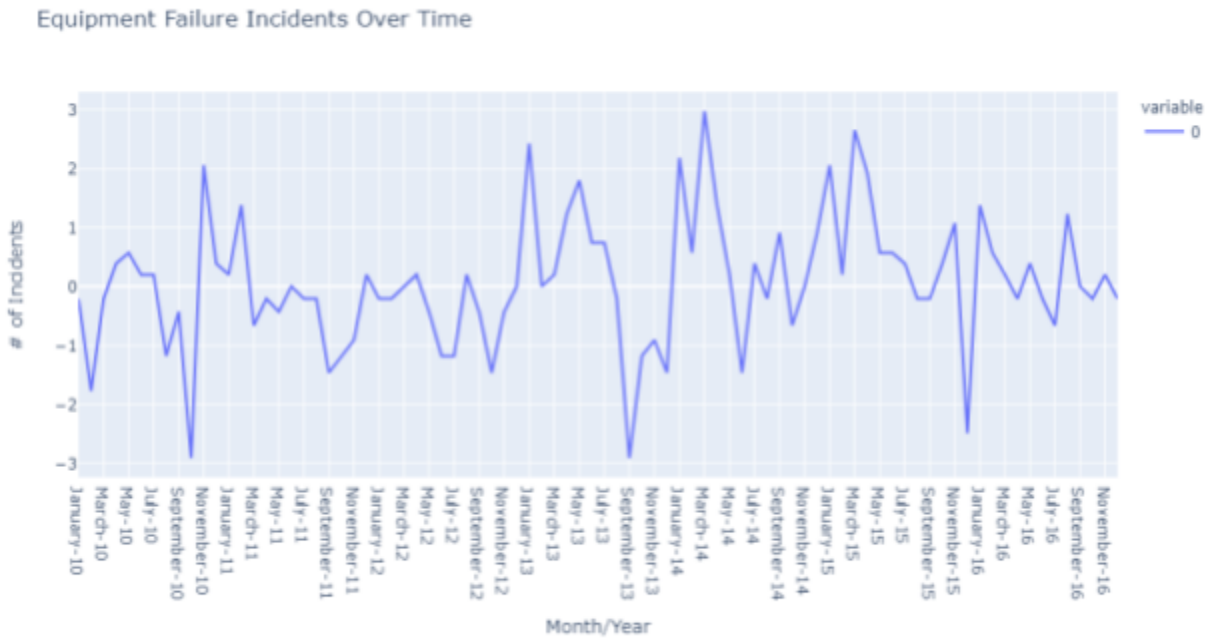


Figure 16: Monthly Raw Time Series Data

Figure 17 shows the seasonality component of the multiplicative decomposition. The graph shows cyclical increases from January to May and decreases from June to September. This is intuitive given the phenomena of thermal contraction/expansion. This phenomena can be mitigated with the implementation of a leak detection system so operators are able to target flange/connection maintenance to specific pipeline locations. The leaks may be gradual to the point where they may be undetectable by traditional SCADA systems until it is too late. In future works, this seasonality effect will be studied on a regional basis to assess if this phenomena is applicable in regions with less severe disparities in seasonal temperature.


Figure 17: Multiplicative Decomposition Seasonal Component

The trend component of the decomposition with seasonality noise removed shows a general upward trend. Referring back to Figure 3 which shows the yearly trend of different cause categories, the lower level of granularity shows a small linear contraction in incidents from 2010 to 2012 followed by linear growth from 2012 to 2014. From 2014 to 2015 the incident rate plateaued and was followed by a contraction to 2016. The multiplicative decomposition offers more granularity on the incident trend. Contrary to the yearly trend, the decomposition shows drastic growth between 2014 to 2015.


Figure 18: Multiplicative Decomposition Trend

The residual distribution shown in Figure 19 resembles the shape of the desired gaussian distribution, but there are deviations that appear non-gaussian. To confirm that the distribution is gaussian, the Shapiro test for normality is applied. The null hypothesis that the distribution is normal is accepted with a p-value > 0.05. When applied to the decomposition residual distribution, the calculated p-value is 0.95, meaning the null hypothesis is accepted. With a gaussian residual distribution, the multiplicative model is a viable descriptive time series model. However, for a forecasting model more sophisticated time series methods will be explored.

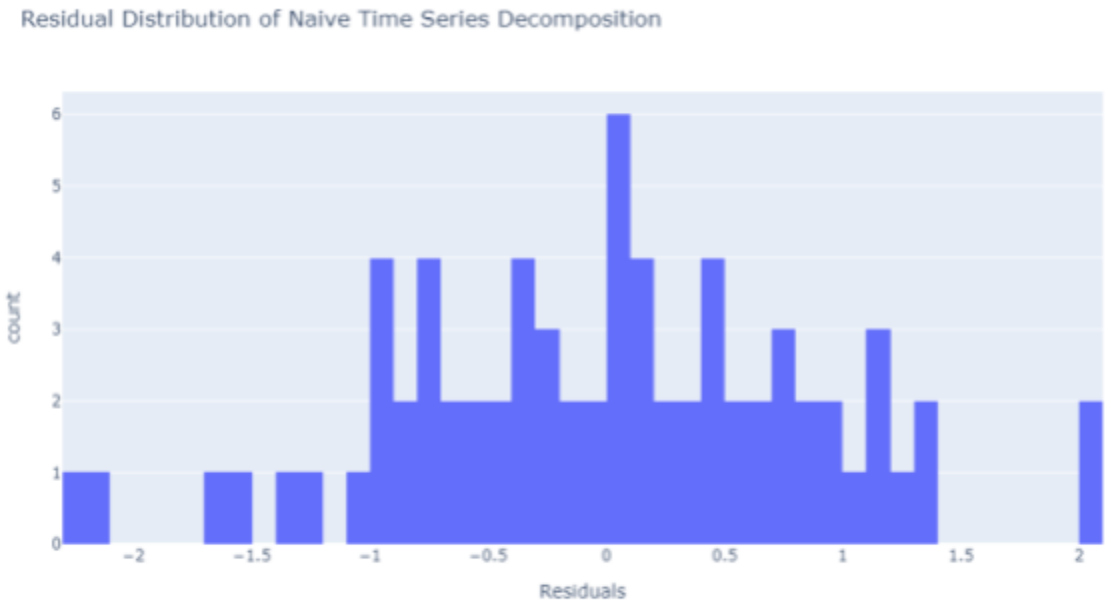


Figure 19: Multiplicative Time Series Residual Distribution

## Forecasting Model Selection

One important aspect to selecting the proper forecasting model is determining if the series is stationary or non-stationary. The multiplicative decomposition model suggests that there is a trend and seasonal component to the time series. To corroborate this assumption the Augmented Dickey Fuller test for non-stationarity is applied. The null hypothesis states that the series data is non-stationary. The resulting p-value is 0.00 meaning that the null

hypothesis can be rejected and the alternative hypothesis of the data being stationary is accepted. This means that there is no significant trend contribution.

An autocorrelation plot of lag features suggests that there is a high correlation at lags 1 and 2. When tuning model hyperparameters, a grid search range of 0 - 3 for the autoregressive component would be an appropriate start.

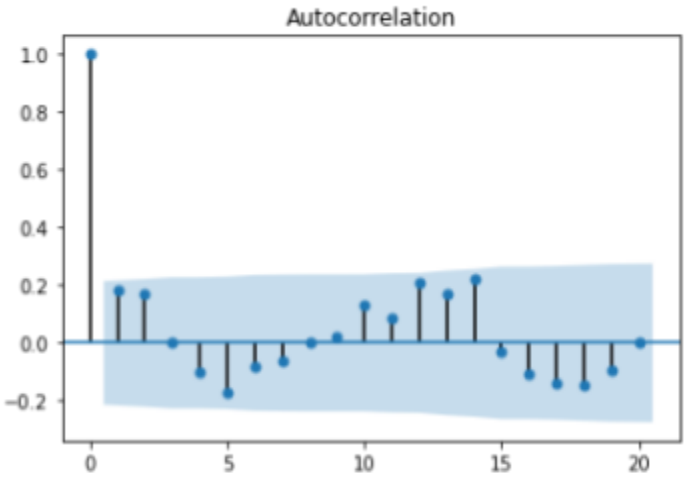


Figure 20: Lag Feature Autocorrelation

## Smoothed Moving Average Predictive Model

With the confirmation that the time series is stationary, the simple smoothed moving average model will be applied. Based on the autocorrelation plot, there is correlation with current time step,t, up to t-2. The model metric being evaluated is Root Mean Squared Error (RMSE). Specifying a window of 3 yielded a RMSE of 1.229, while a window of 2 yielded a RMSE of 1.234.

Figure 21 shows the predictions of the smoothed moving average model relative to the actual values. The final predicted value differed from the expected result by 0.2. The next evaluation of the model will be the distribution of residuals and autocorrelation of residual lag features.

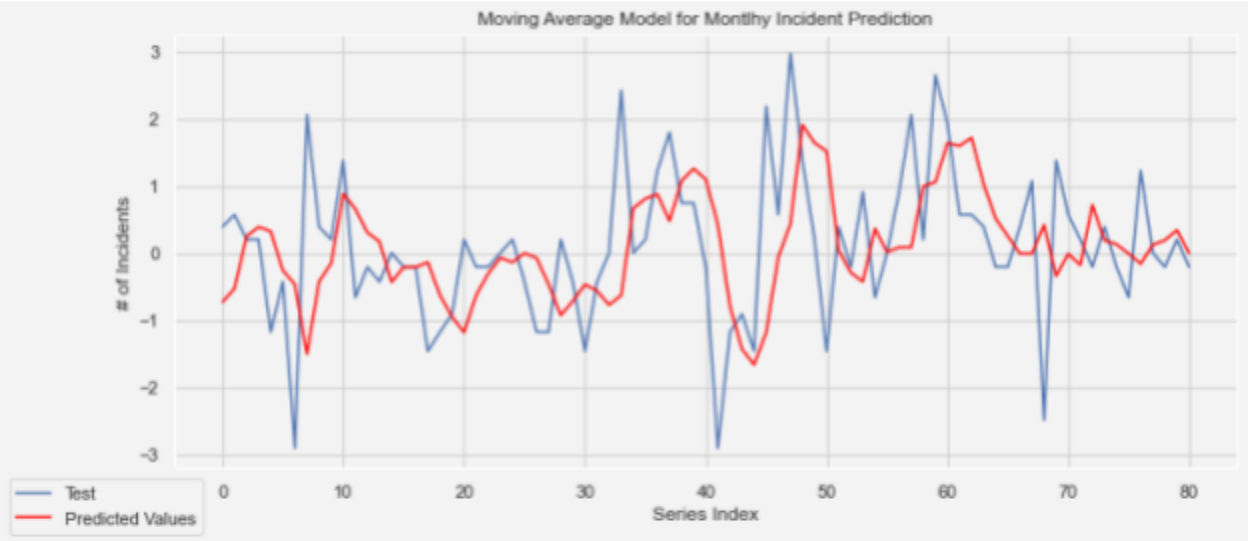


Figure 21: Smoothed Moving Average Predictive Model

The residual distribution shown in Figure 22 resembles the shape of a gaussian distribution. The Shapiro test for normality resulted in a p-value of 0.016, meaning the null hypothesis of normality is rejected. More sophisticated methods must be used to meet the normality assumption.

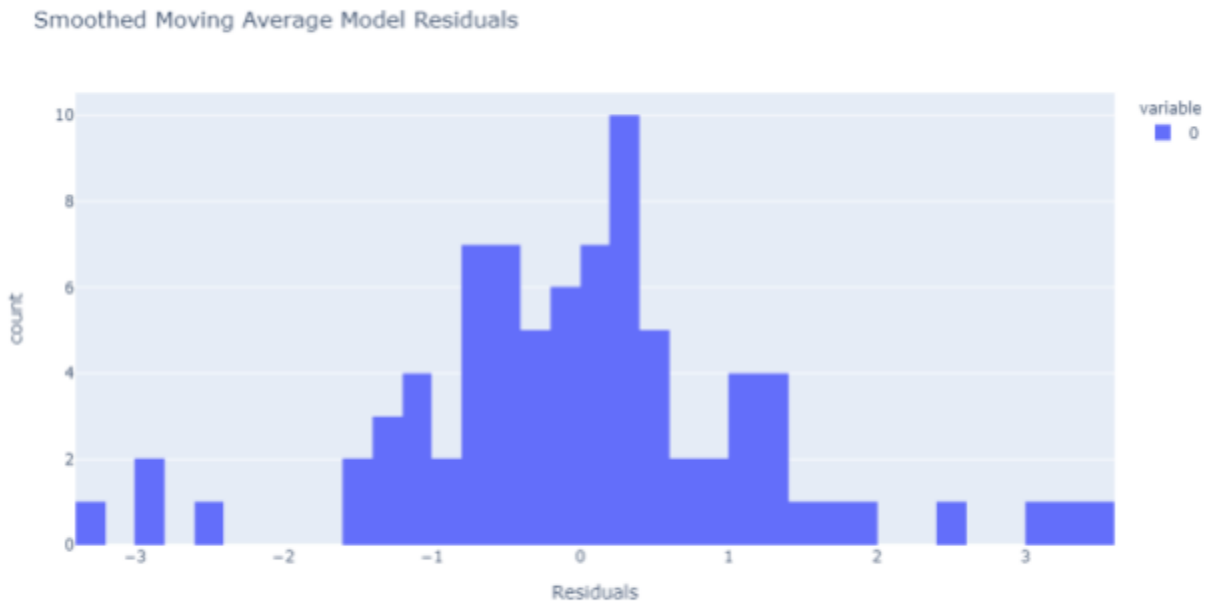


Figure 22: Residual Distribution of Smoothed Moving Average Model

## SARIMAX Model Tuning

SARIMAX is a variation of ARIMA that contains a seasonal order component. A manual hyperparameter grid search was performed using the method of walk-forward validation for model training. The grid search varies the following variables;

- Non-Seasonal Order
    - p: Autoregressive Variable
    - d: Differencing
        - Note: Differencing was used to observe the effect on RMSE. The ADF test of non-stationarity suggests negligible trend effects
    - q: Moving Average Parameters
        - Moving average of the residual errors
- Seasonal Order
    - P: Autoregressive Variable
    - D: Differencing
    - Q: Moving Average Parameters
    - s: Periodicity (constant = 12)
- Trend
    - Despite showing no significant trend effects, the following trend parameters were varied
        - No Trend
        - Constant
        - Linear Trend
        - Constant Linear Trend

Performing grid search walk forward validation yielded a model with a minimized RMSE of 0.930 with an order of (0,0,0), seasonal order of (2,0,1,12), and a linear trend.

Figure 23 shows the predicted and test values of the model with the lowest RMSE. The SARIMAX model does not appear to fit the shape of the test set as well as the smoothed moving average, but has a 0.3 improvement in terms of RMSE.

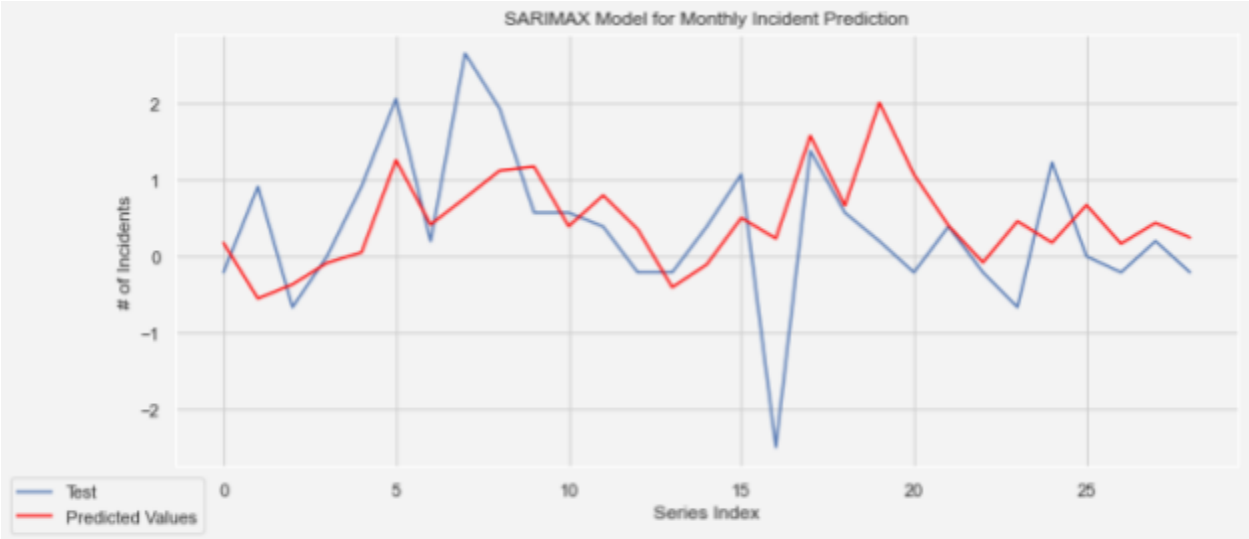


Figure 23: SARIMAX Test and Predicted Results

The SARIMAX model residual distribution is shown in Figure 24. The Shapiro test of normality yields a p-value of 0.27, meaning that the null hypothesis of normality is accepted.

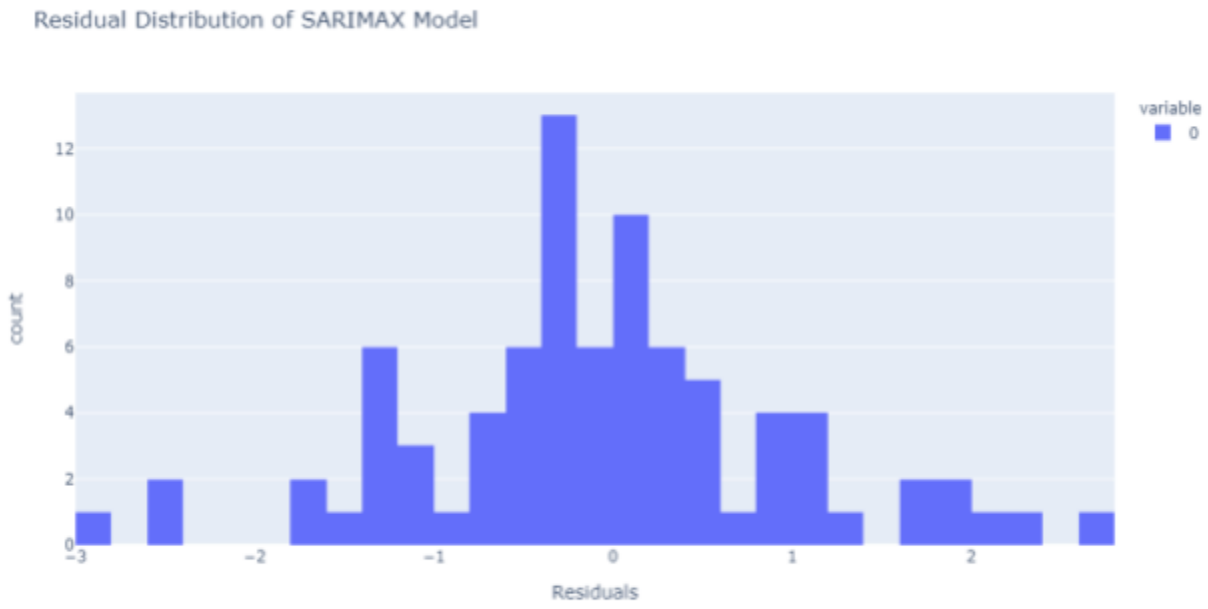


Figure 24: Residual Distribution of SARIMAX Model

# Conclusion

At the yearly level, the trend in pipeline incidents has increased significantly over the time period of 2010 - 2016. The most frequent causes of pipeline incidents are equipment/weld/material failure, corrosion, and operator error. A majority of incidents that occur are minor with low downtime, net product loss, and costs beneath $200,000. However, there is a high degree of outlier effect in the cost distribution. Regions with more severe winter conditions have higher costs per incident due to thermal contraction/expansion. Operators with lower volume capacity assets were more frequent on the top 20 operators in terms of cost per incident. This is likely caused by lower technology SCADA/reliability systems. When selecting a time series model for forecasting monthly equipment failure incidents, the model that minimized RMSE was a SARIMAX model with a seasonal order of (2,0,1,12) with a linear trend component.

# Appendix

## Data Columns

['Report Number', 'Supplemental Number', 'Accident Year', 'Operator ID',

'Operator Name', 'Pipeline/Facility Name', 'Pipeline Location',

'Pipeline Type', 'Liquid Type', 'Liquid Subtype', 'Liquid Name',

'Accident City', 'Accident County', 'Accident State',

'Accident Latitude', 'Accident Longitude', 'Cause Category',

'Cause Subcategory', 'Unintentional Release (Barrels)',

'Intentional Release (Barrels)', 'Liquid Recovery (Barrels)',

'Net Loss (Barrels)', 'Liquid Ignition', 'Liquid Explosion',

'Pipeline Shutdown', 'Shutdown Date/Time', 'Restart Date/Time',

'Public Evacuations', 'Operator Employee Injuries',

'Operator Contractor Injuries', 'Emergency Responder Injuries',

'Other Injuries', 'Public Injuries', 'All Injuries',

'Operator Employee Fatalities', 'Operator Contractor Fatalities',

'Emergency Responder Fatalities', 'Other Fatalities',

'Public Fatalities', 'All Fatalities', 'Property Damage Costs',

'Lost Commodity Costs', 'Public/Private Property Damage Costs',

'Emergency Response Costs', 'Environmental Remediation Costs',
'Other Costs', 'All Costs', 'Down Time', 'Region']

## Region Definition


Figure 25: U.S Regions Defined by the Petroleum Administration

## Works Cited

[1] U.S Energy Information Administration