Project 5 | Group 1
Jack Vaughan
Ella Shafi
Kathy Kaviani
Troy Alexander

Retail Sales Forecasting

WoMart Stores

Our team

Jack Vaughan

ChatGPT in human form

Ella Shafi

Did not like Everything Everywhere All at Once

Kathy Kaviani

Has never answered a GA ice breaker

Troy Alexander

Accidentally solved a rubik's cube

Table of contents

01

Problem / Objective

What problem are we solving for?

Problem Statement

Antiquated forecasting

Historically reactionary and unable to accurately forecast sales and demand.

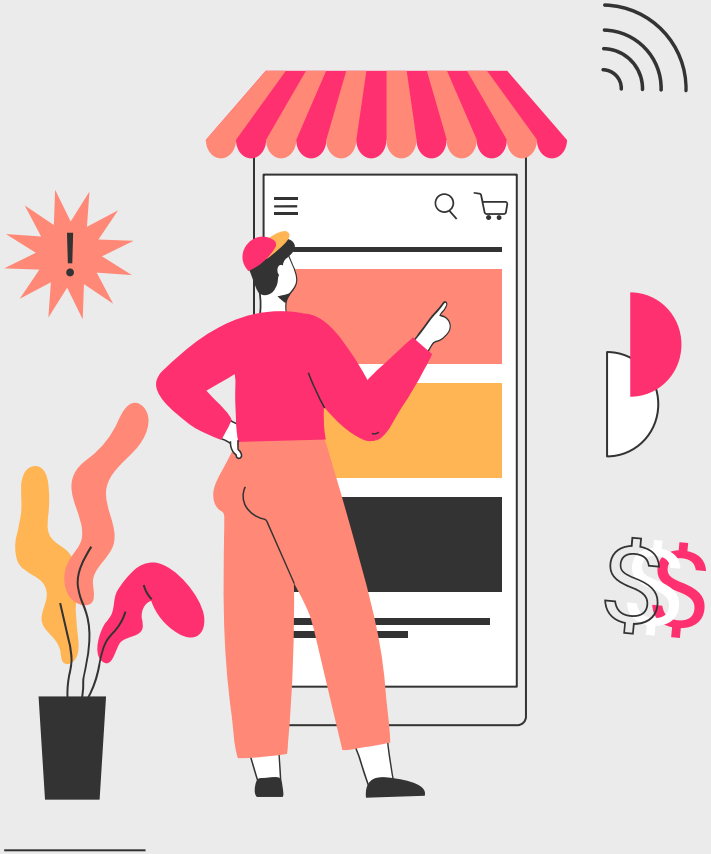
Increasing labor costs

Inflation in labor costs and labor supply has negatively affected profitability.

Workforce optimization

With no ability to accurately forecast demand, WoMart is unable to efficiently deploy workforce.

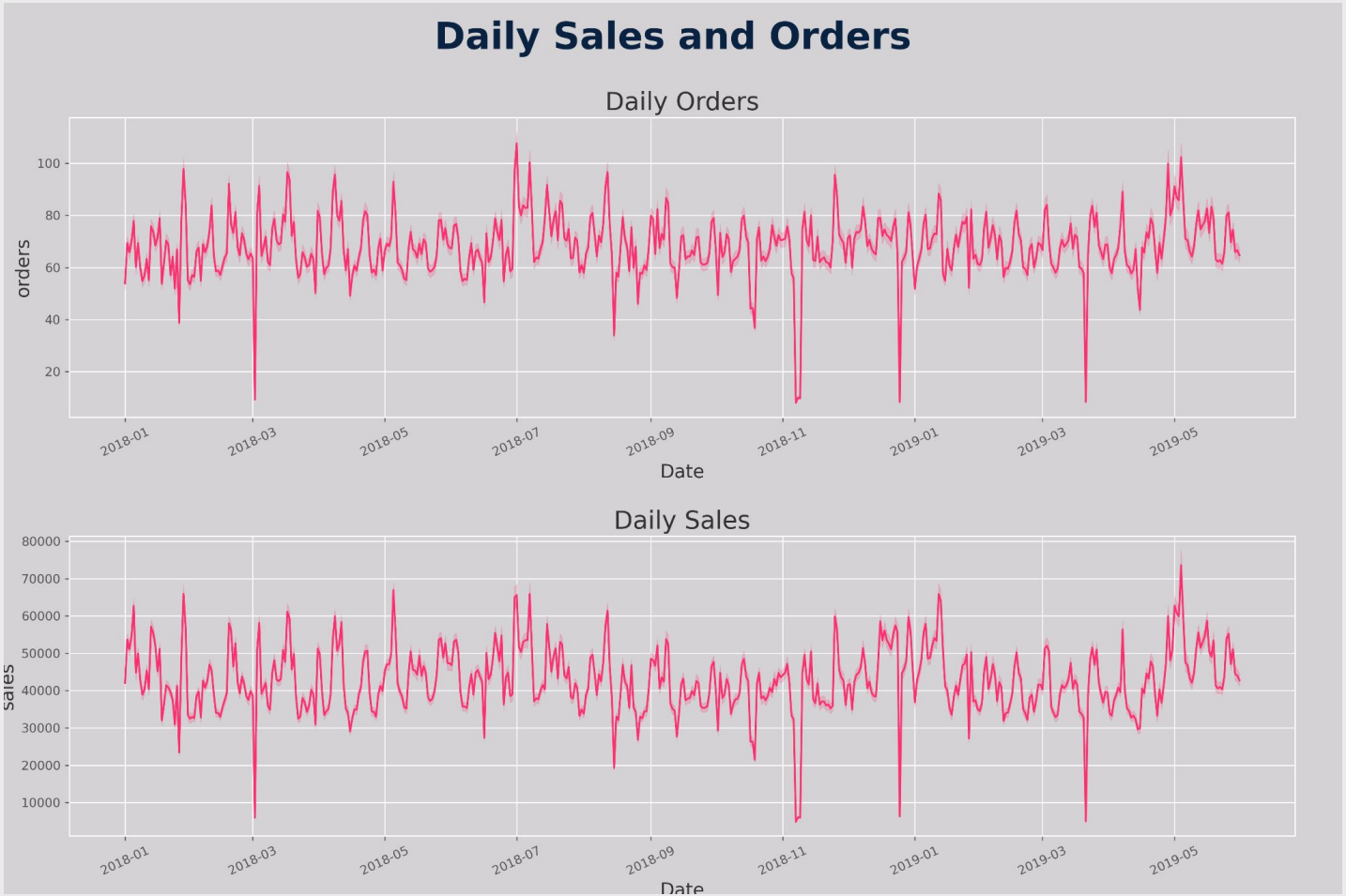
Our goal is to **forecast sales by region** for the following month in order to assist in **workforce optimization**.



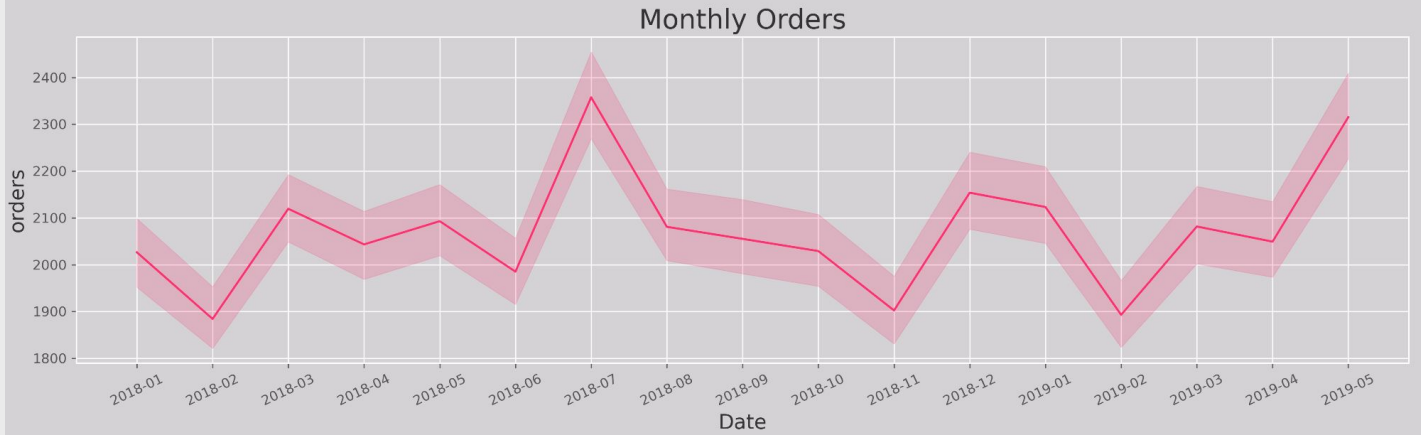
02

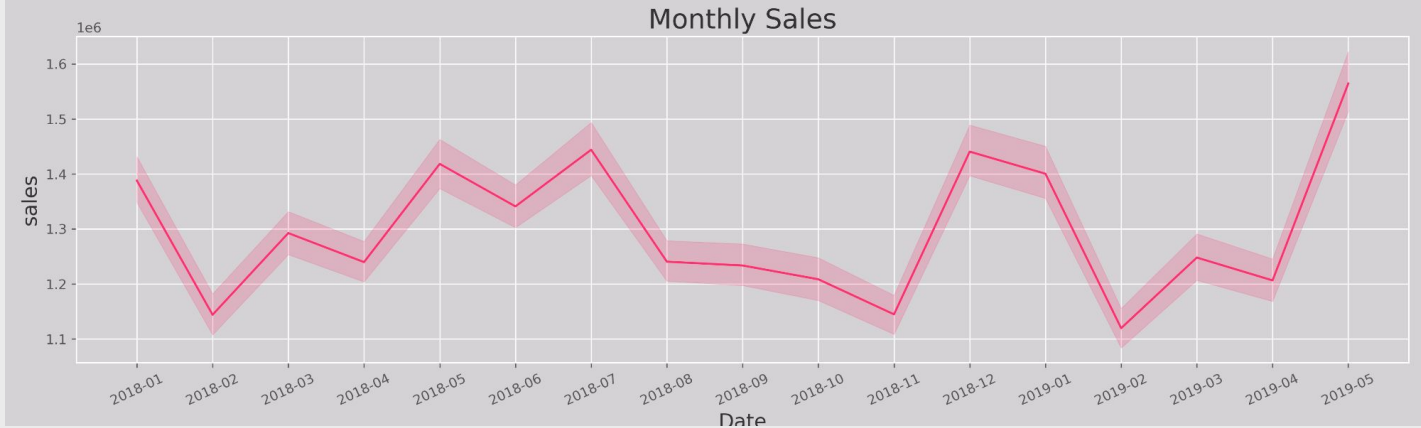
Background

Exploratory data analysis and basis of presentation

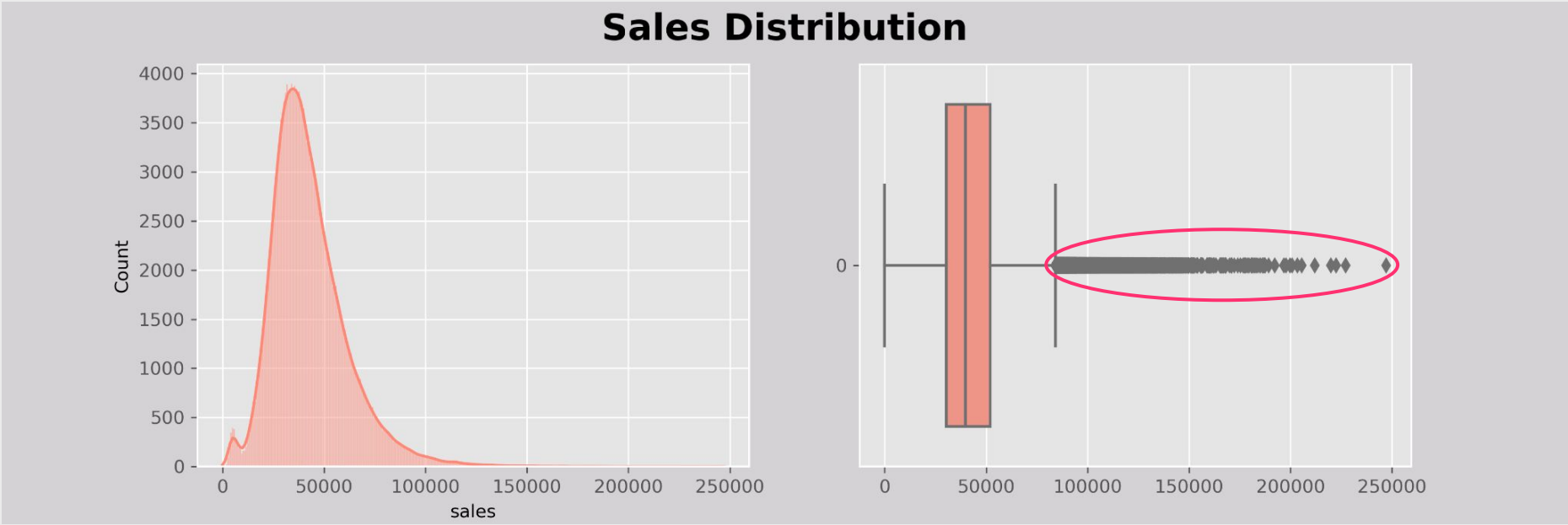

Daily Sales and Orders
Daily Orders
orders
100
80
60
40
20
2018-01
2018-03
2018-05
2018-07
2018-09
2018-11
2019-01
2019-03
2019-05
Date
Daily Sales
sales
80000
70000
60000
50000
40000
30000
20000
10000
2018-01
2018-03
2018-05
2018-07
2018-09
2018-11
2019-01
2019-03
2019-05
Date

Monthly Sales and Orders





DAILY SALES DISTRIBUTION



- Sales over 100K belong to the same region(R2), Store_type (S4), Location_type(L2)

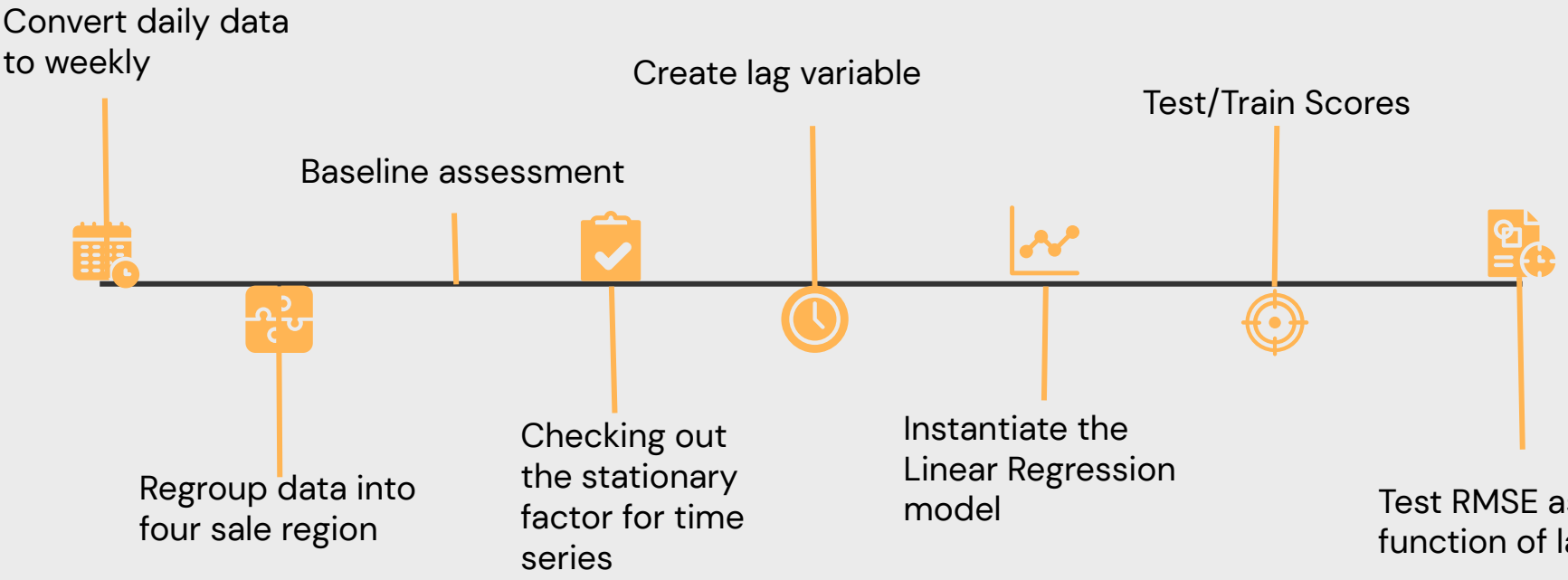
SALES AVERAGE BY STORE ATTRIBUTES

03

Forecast Model

Production model and evaluation metrics

OUR MODELING PATH

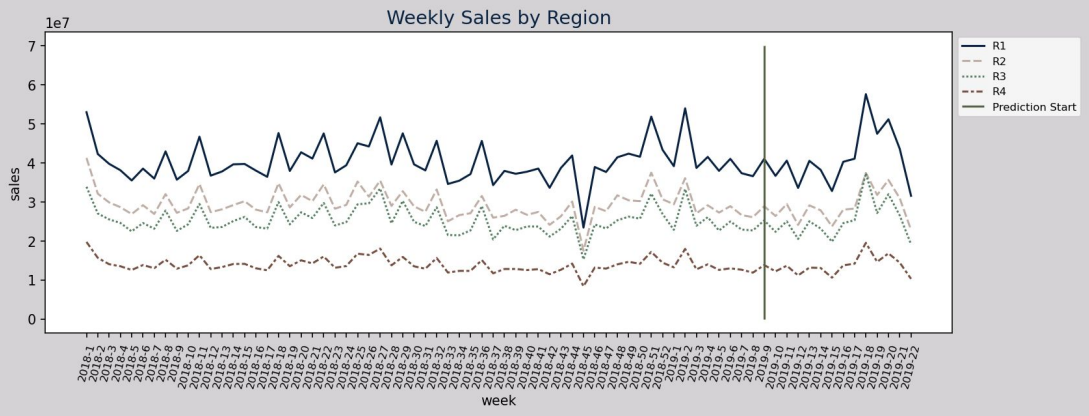


Stationary Check

Significance level: 0.05

P-value : 0.008

Confidence: 99.2%



Stationarity refers to time series that have relatively constant statistical values -- think mean, variance, etc.

Necessary to compute because stationarity is an assumption of modeling time series data.

P-value gives us ~99% confidence in stationarity.

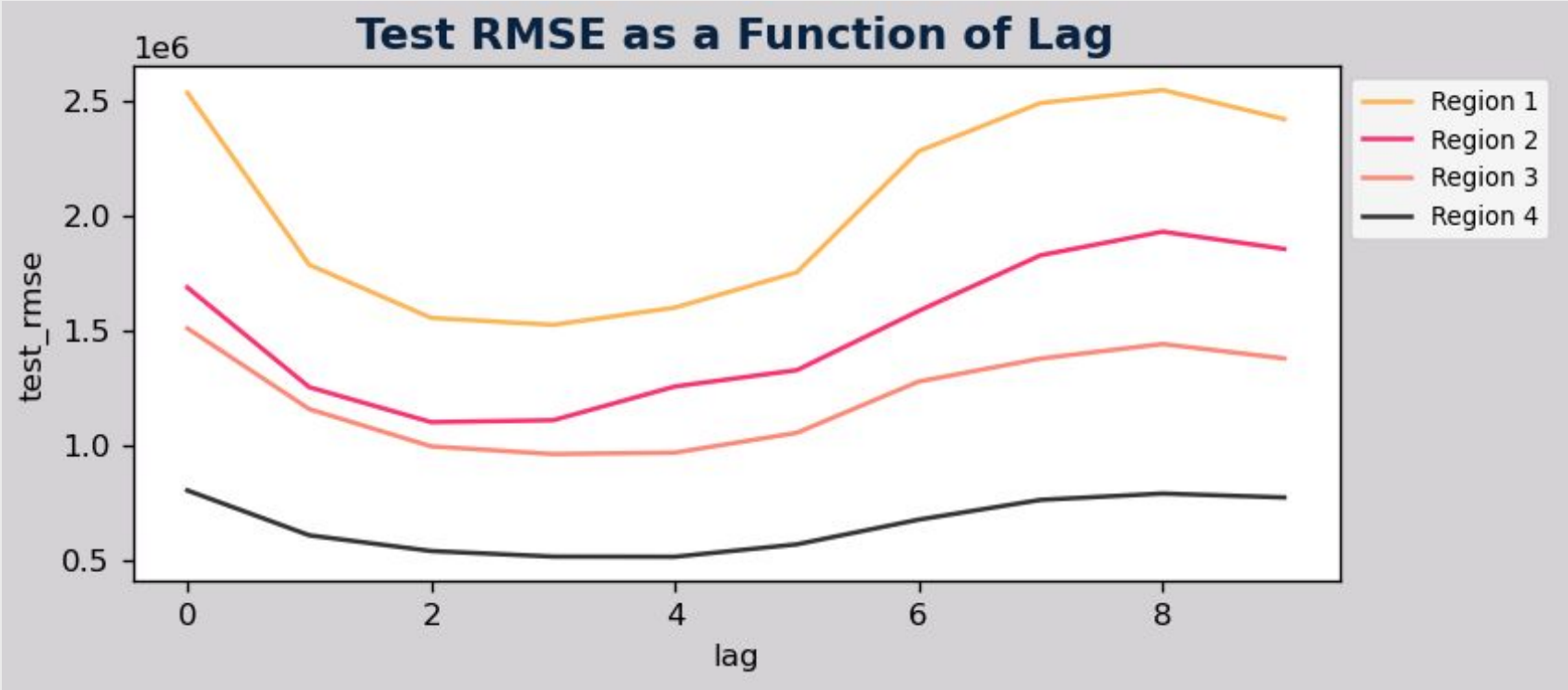
What is Lag?

- Variable lagging is a common practice in time series analysis. The idea is to add "lagged" versions of the dependent variable, in this case the sales data, to the model.

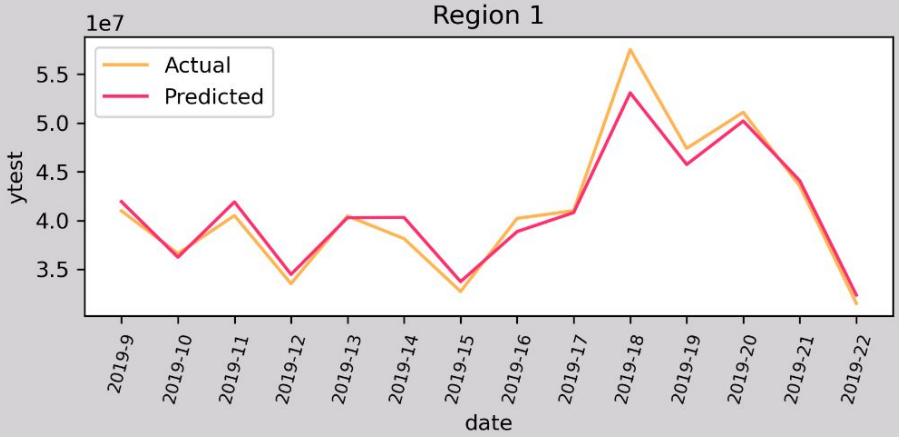
- We accomplish this by creating a new column for each lagged variable, and then shifting the data in the column by the number of periods we want to lag the variable.

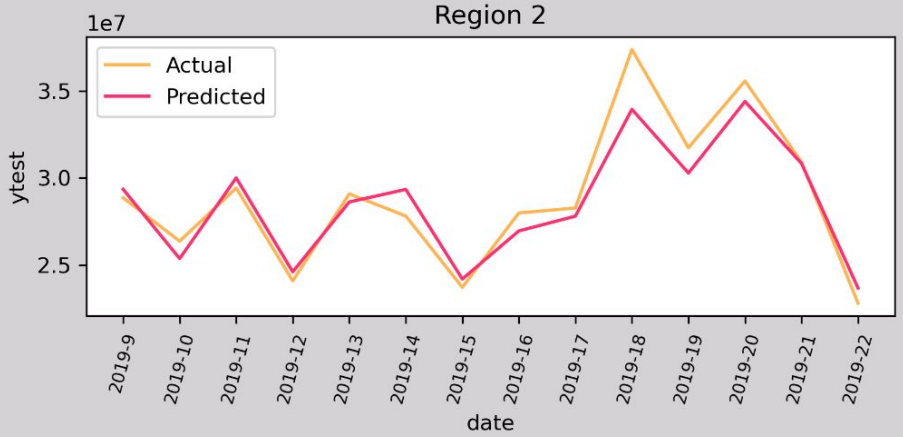
- Since our data is separated into regions, we will need to separate the data into regions and then lag the data for each region separately.

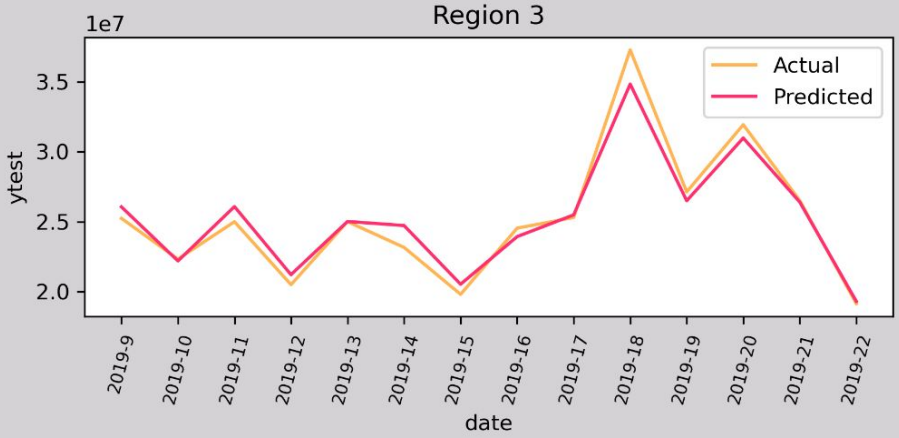
sales-1	sales-2	sales-3	sales-4	sales-5
NaN	NaN	NaN	NaN	NaN
52925573.22	NaN	NaN	NaN	NaN
42225337.95	52925573.22	NaN	NaN	NaN
39719301.00	42225337.95	52925573.22	NaN	NaN
38068777.05	39719301.00	42225337.95	52925573.22	NaN

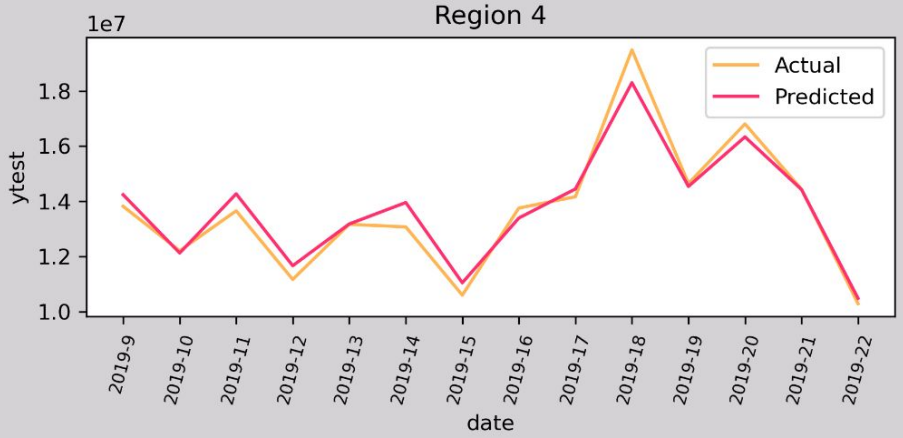

Test RMSE as a Function of Lag
1e6
2.5
2.0
1.5
1.0
0.5
test_rmse
0
2
4
6
8
lag
Region 1
Region 2
Region 3
Region 4

RESULT


Region 1
1e7
Actual
Predicted
ytest
5.5
5.0
4.5
4.0
3.5
2019-9
2019-10
2019-11
2019-12
2019-13
2019-14
2019-15
2019-16
2019-17
2019-18
2019-19
2019-20
2019-21
2019-22
date


Region 2
1e7
Actual
Predicted
ytest
3.5
3.0
2.5
2019-9
2019-10
2019-11
2019-12
2019-13
2019-14
2019-15
2019-16
2019-17
2019-18
2019-19
2019-20
2019-21
2019-22
date


Region 3
1e7
Actual
Predicted
ytest
3.5
3.0
2.5
2.0
2019-9
2019-10
2019-11
2019-12
2019-13
2019-14
2019-15
2019-16
2019-17
2019-18
2019-19
2019-20
2019-21
2019-22
date


Region 4
1e7
Actual
Predicted
ytest
1.8
1.6
1.4
1.2
1.0
2019-9
2019-10
2019-11
2019-12
2019-13
2019-14
2019-15
2019-16
2019-17
2019-18
2019-19
2019-20
2019-21
2019-22
date

04

Conclusion

Recommendations and outcome of analysis

Conclusion

- WoMart should deploy sales forecasting model to drive real-time decisions regarding workforce optimization.

Model Scores

Region	Train_R2	Test_R2	Train_RMSE	Test_RMSE
R1	0.7699	0.9461	2.4e+06	1.59e+06
R2	0.7602	0.9021	1.64e+06	1.25e+06
R3	0.7922	0.9560	1.49e+06	9.66e+05
R4	0.7725	0.9510	8.17e+05	5.13e+05

Conclusion

- Evaluate the results of the model to understand the impact of discounts and holidays to optimize sales and orders.

Discount	Average sales
0	$37403
1	$49426

Model Recommendations

With additional budget for development team...

- Increase historical data for more accurate training
- Further parameter tuning
- Consideration of other deep learning time series models (RNN, ARIMA)
- Incorporate profitability into model to understand cost/benefit tradeoffs
- Investigate more features that may affect sales

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution