Project
Capstone

SALES FORECASTING PROJECT

Presented by: Nurhidayah
Adviced by: Dr Nor Azizah Ali

CONTENT

PROBLEM STATEMENT

The business problem revolves around improving profitability and operational efficiency through better insights into product sales and profit trends. Two key business insight to be focused on are profitability analysis and sales forecasting.

First, profitability analysis is to identify patterns of profit loss, both geographically and by product category, is critical for the sales department to understand underperforming areas and products. This will help in making strategic decisions regarding product focus, marketing efforts, and regional investments.

Second, sales forecasting is to estimate future sales for January 2022. This will help in planning inventory, marketing, and resource allocation. Accurate forecasting ensures that

DATA DESCRIPTION

Provided with 3 Excel files (Orders_2017.xlsx, Orders_2018.xlsx, Orders_2019.xlsx) and 2 CSV files (Orders_2020.csv, Orders_2021.csv) that contain information about order date, customer, region, product category, sub-category, sales, profit, shipping cost, and return status.

SAMPLE DATA

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Category	City	Country/Region	Discount	Order ID	Postal Code	Product Name	Profit	Quantity	Region	Sales	Segment	Ship Mode	State	Sub-Category	OrderDate	ShipDate	Days to Ship	Scheduled
2	Furniture	Fort Lauder	United States	45	US-2017-108	33311	Bretford CR4500 Series Slim F	-383	5	South	958	Consumer	Standard Class	Florida	Tables	11/10/2017	18/10/2017	6	
3	Office Supplies	Fort Lauder	United States	20	US-2017-108	33311	Eldon Fold 'N Roll Cart Systen	3	2	South	22	Consumer	Standard Class	Florida	Storage	11/10/2017	18/10/2017	6	
4	Office Supplies	Fort Worth	United States	80	US-2017-118	76106	Holmes Replacement Filter fo	-124	5	Central	69	Home Office	Standard Class	Texas	Appliances	22/11/2017	26/11/2017	6	
5	Office Supplies	Fort Worth	United States	80	US-2017-118	76106	Storex DuraTech Recycled Pla	-4	3	Central	3	Home Office	Standard Class	Texas	Binders	22/11/2017	26/11/2017	6	
6	Furniture	Orem	United States	0	CA-2017-106	84057	Bretford CR4500 Series Slim F	240	3	West	1045	Consumer	Standard Class	Utah	Tables	25/9/2017	30/9/2017	6	

```
1  Row ID,Order ID,Order Date,Ship Date,Ship Mode,Customer Name,Segment,Country/Region,City,State,Postal Code,Region,Category,Sub-Category,Product Name,Sales,Quantity,Discount,Profit,Days to Sh
2  1,CA-2020-152156,8/11/2020 0:00,11/11/2020 0:00,Second Class,Claire Gute,Consumer,United States,Henderson,Kentucky,42420,South,Furniture,Bookcases,Bush Somerset Collection Bookcase,261.96,2,
3  2,CA-2020-152156,8/11/2020 0:00,11/11/2020 0:00,Second Class,Claire Gute,Consumer,United States,Henderson,Kentucky,42420,South,Furniture,Chairs,"Hon Deluxe Fabric Upholstered Stacking Chairs
4  3,CA-2020-138688,12/6/2020 0:00,16/6/2020 0:00,Second Class,Darrin Van Huff,Corporate,United States,Los Angeles,California,90036,West,Office Supplies,Labels,Self-Adhesive Address Labels for
5  14,CA-2020-161389,5/12/2020 0:00,10/12/2020 0:00,Standard Class,Irene Maddox,Consumer,United States,Seattle,Washington,98103,West,Office Supplies,Binders,Fellowes PB200 Plastic Comb Binding
6  22,CA-2020-137330,9/12/2020 0:00,13/12/2020 0:00,Standard Class,Ken Black,Corporate,United States,Fremont,Nebraska,68025,Central,Office Supplies,Art,Newell 318,19.46,7,0,5.0596,6
7  23,CA-2020-137330,9/12/2020 0:00,13/12/2020 0:00,Standard Class,Ken Black,Corporate,United States,Fremont,Nebraska,68025,Central,Office Supplies,Appliances,"Acco Six-Outlet Power Strip, 4' C
8  26,CA-2020-121755,16/1/2020 0:00,20/1/2020 0:00,Second Class,Eric Hoffmann,Consumer,United States,Los Angeles,California,90049,West,Office Supplies,Binders,Wilson Jones Active Use Binders,11
9  27,CA-2020-121755,16/1/2020 0:00,20/1/2020 0:00,Second Class,Eric Hoffmann,Consumer,United States,Los Angeles,California,90049,West,Technology,Accessories,Imation 8GB Mini TravelDrive USB 2.
10 36,CA-2020-117590,8/12/2020 0:00,10/12/2020 0:00,First Class,Gene Hale,Corporate,United States,Richardson,Texas,75080,Central,Technology,Phones,GE 30524EE4,1097.544,7,0.2,123.4737,1
11 37,CA-2020-117590,8/12/2020 0:00,10/12/2020 0:00,First Class,Gene Hale,Corporate,United States,Richardson,Texas,75080,Central,Furniture,Furnishings,"Electrix Architect's Clamp-On Swing Arm L
12 43,CA-2020-101343,17/7/2020 0:00,22/7/2020 0:00,Standard Class,Ruben Ausman,Corporate,United States,Los Angeles,California,90049,West,Office Supplies,Storage,"Eldon Base for stackable storag
```

DATA PREPROCESSING

DATA PREPROCESSING

Import Dataset

```
#import libraries
import gdown
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import chardet
```

```
# download data from drive
shared_url = 'https://drive.google.com/file/d/1y35PeUIW-BO2RDqYOzp-o9g7ddWdnKXZ/view?usp=sharing'
file_url    = f'https://drive.google.com/uc?id={shared_url.split("/")[-2]}'
output = 'ORDERS_2021.csv'
gdown.download(file_url, output, quiet=False)

shared_url = 'https://drive.google.com/file/d/1dRjlYrNvHknNIvT6eWNhe9w0_9GU6YXN/view?usp=sharing'
file_url    = f'https://drive.google.com/uc?id={shared_url.split("/")[-2]}'
output = 'ORDERS_2020.csv'
gdown.download(file_url, output, quiet=False)

shared_url = 'https://docs.google.com/spreadsheets/d/1yn7J3i5gVfYEZEDXCVbOT8s-ITyLiBzN/edit?usp=sharing&ouid=115467567108565679378&rtpof=true&sd=true'
file_url    = f'https://drive.google.com/uc?id={shared_url.split("/")[-2]}'
output = 'ORDER_2019.xlsx'
gdown.download(file_url, output, quiet=False)

shared_url = 'https://docs.google.com/spreadsheets/d/19AjIVfQWU6i8Qlif_EUkt3mi0BNvlBO1/edit?usp=sharing&ouid=115467567108565679378&rtpof=true&sd=true'
file_url    = f'https://drive.google.com/uc?id={shared_url.split("/")[-2]}'
output = 'ORDER_2018.xlsx'
gdown.download(file_url, output, quiet=False)

```

DATA PREPROCESSING

Explore Dataset

```
# add ORDER_2018 data into dataframe
df_order2018 = pd.read_excel('ORDER_2018.xlsx')
df_order2018.head()
```

	Category	City	Country/Region	Discount	Order ID	Postal Code	Product Name	Profit	Quantity	Region	Sales	Segment	Ship Mode	State	Sub-Category	OrderDate	ShipDate	Days to Ship Scheduled
0	Furniture	Henderson	United States	0	CA-2018-152156	42420.0	Bush Somerset Collection Bookcase	42	2	South	262	Consumer	Second Class	Kentucky	Bookcases	2018-11-08	2018-11-11	3.0
1	Furniture	Henderson	United States	0	CA-2018-152156	42420.0	Hon Deluxe Fabric Upholstered Stacking Chairs,...	220	3	South	732	Consumer	Second Class	Kentucky	Chairs	2018-11-08	2018-11-11	3.0
2	Office Supplies	Los Angeles	United States	0	CA-2018-138688	90036.0	Self-Adhesive Address Labels for Typewriters b...	7	2	West	15	Corporate	Second Class	California	Labels	2018-06-12	2018-06-16	3.0
3	Office Supplies	Seattle	United States	20	CA-2018-161389	98103.0	Fellowes PB200 Plastic Comb Binding Machine	133	3	West	408	Consumer	Standard Class	Washington	Binders	2018-12-05	2018-12-10	6.0

```
# check entries and column
df_order2021.shape, df_order2020.shape, df_order2019.shape, df_order2018.shape, df_order2017.shape
```

```
((3312, 20), (2587, 20), (4278, 18), (3179, 18), (2522, 18))
```

DATA PREPROCESSING

Data Filtration

```
# Drop extra columns
df_order2021.drop(['Row ID', 'Customer Name'], axis=1, inplace=True)
df_order2020.drop(['Row ID', 'Customer Name'], axis=1, inplace=True)

df_order2021.shape, df_order2020.shape
```

DATA PREPROCESSING

Data Cleaning

```python
# rename columns in df_order2019, df_order2018, and df_order2017
dfList = [df_order2017,df_order2018,df_order2019]
for df in dfList:
  df.rename(columns={'OrderDate': 'Order Date', 'ShipDate':'Ship Date'},
  inplace=True)
  print(df.columns)
```

```python
# Convert datatype object to datetime for columns in df_order2021 and
df_order2020
dfList = [df_order2021,df_order2020]
for df in dfList:
  df['Order Date'] = pd.to_datetime(df['Order Date'], format="%d/%m/%Y %H:%M")
  df['Ship Date'] = pd.to_datetime(df['Ship Date'], format="%d/%m/%Y %H:%M")
  print(df['Order Date'].dtypes)
  print(df['Ship Date'].dtypes)
```

DATA PREPROCESSING

Export

```
#save into excel
order_return_df.to_excel('order_return_df_3.xlsx', index=False)
```

EXPLORATORY DATA ANALYSIS

```
# Descriptive Statistics for Numerical Columns
order_return_df.describe()
```

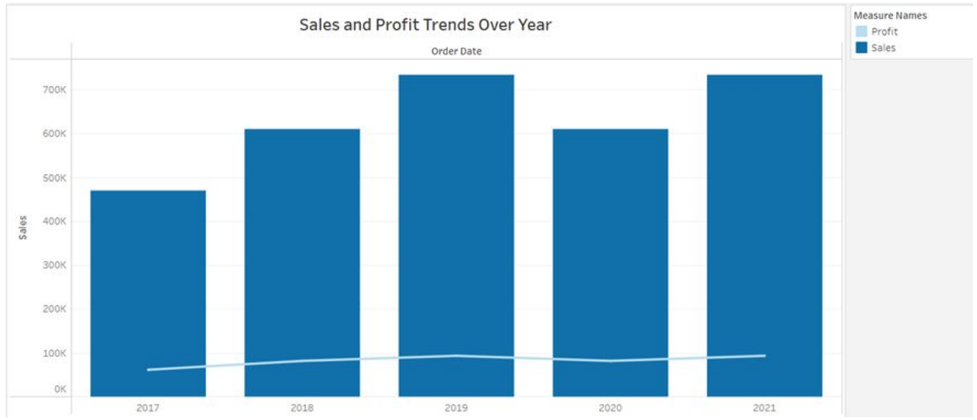
	Order Date	Ship Date	Sales	Quantity	Discount	Profit	Days to Ship Scheduled
count	13900	13900	13900.000000	13900.000000	13900.000000	13900.000000	13900.000000
mean	2019-10-11 14:48:20.719424512	2019-10-15 13:19:39.971222784	227.015817	3.784532	9.027449	29.644708	4.491649
min	2017-01-02 00:00:00	2017-01-04 00:00:00	0.000000	1.000000	0.000000	-6600.000000	1.000000
25%	2018-09-04 00:00:00	2018-09-07 00:00:00	17.380000	2.000000	0.000000	2.000000	3.000000
50%	2019-10-19 00:00:00	2019-10-22 00:00:00	54.686000	3.000000	0.200000	8.750700	6.000000
75%	2020-12-13 00:00:00	2020-12-18 00:00:00	207.036000	5.000000	20.000000	29.000000	6.000000
max	2021-12-30 00:00:00	2022-01-05 00:00:00	17500.000000	14.000000	80.000000	8400.000000	6.000000
std	NaN	NaN	602.037902	2.217501	17.430890	249.769539	1.948143

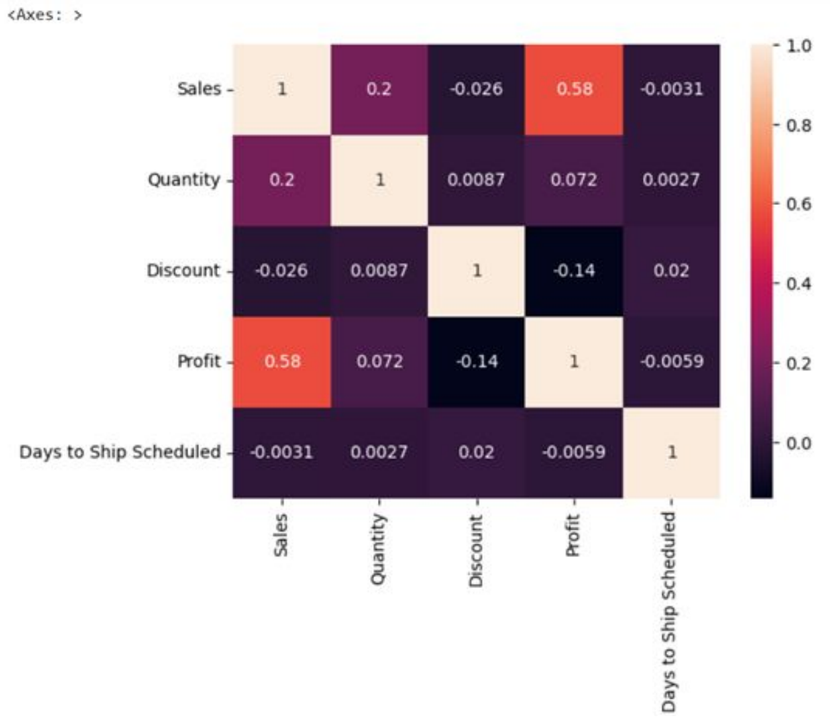
```
[40]  # Value counts for categorical variables
      order_return_df['Region'].value_counts()
```

Region	count
West	4442
East	4018
Central	3238
South	2202

dtype: int64

EXPLORATORY DATA ANALYSIS





EXPLORATORY DATA ANALYSIS

MODEL BUILDING

- Phase where predictive models are constructed, trained, and assessed for their performance.
- Metrics such as accuracy, precision, recall, and F1-score are employed to measure model performance, and techniques like cross-validation help estimate generalization.
- The goal is to build models that provide actionable insights or predictions, ensuring their effectiveness and reliability for real-world applications.

MODEL BUILDING

ARIMA

ARIMA stands for AutoRegressive Integrated Moving Average, a statistical time series model commonly used for forecasting. It's popular in cases where data shows patterns over time and can be decomposed into components such as trend, seasonality, and noise. ARIMA models are especially effective for series that are non-stationary but can be made stationary through differencing.

```python
from statsmodels.tsa.arima.model import ARIMA

# fit model
model_arima = ARIMA (daily_sales_final['Sales'], order = (1,1,0))
model_fit_arima = model_arima.fit()

# summary of fit model
print(model_fit_arima.summary())
```

```python
# split into train and test sets
X = daily_sales_final['Sales']
size = int(len(X) * 0.80)
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]


# walk-forward validation
prediction_arima = test.copy()
for t in range(len(test)):
    model = ARIMA (history, order = (1,1,0))
    model_fit = model.fit()
    output = model_fit.forecast()
    yhat = output[0]
    prediction_arima.iloc[t] = yhat # Use iloc to assign by position
    obs = test.iloc[t] # Use iloc to access by position
    history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))
```

MODEL BUILDING

ETS

The ETS model is a part of Exponential Smoothing methods, which assign exponentially decreasing weights to past observations. The ETS framework is particularly good for modeling time series that exhibit trends and seasonality. It automatically identifies the nature of the trend and seasonality components to produce a forecast.

```python
# fit an ETS model and plot residual errors
from statsmodels.tsa.holtwinters import ExponentialSmoothing

# fit model
model_es = ExponentialSmoothing(daily_sales_final['Sales'],
                                trend='add',
                                seasonal='add',
                                seasonal_periods=365)
model_fit_es = model_es.fit()
# summary of fit model
print(model_fit_es.summary())
```

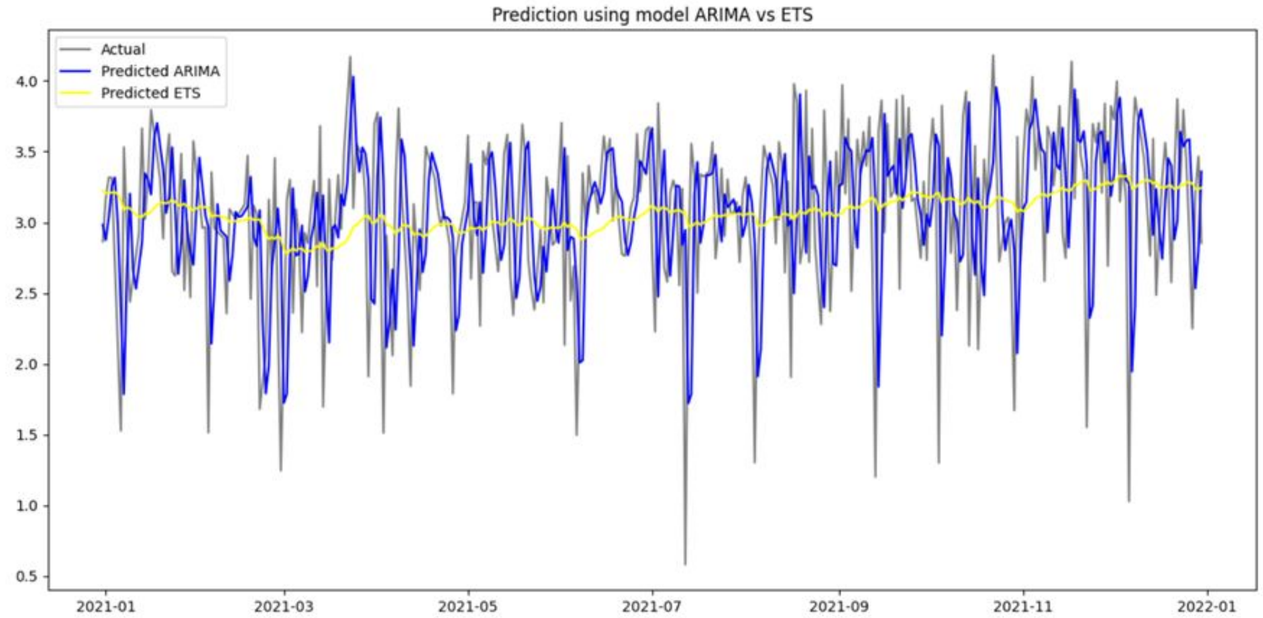
```python
# split into train and test sets
X = daily_sales_final['Sales']
size = int(len(X) * 0.8)
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]

# walk-forward validation
prediction_es = test.copy()
for t in range(len(test)):
  model = ExponentialSmoothing (history)
  model_fit = model.fit()
  output = model_fit.forecast()
  yhat = output[0]
  prediction_es.iloc[t] = yhat
  obs = test.iloc[t]
  history.append(obs)
  print('predicted=%f, expected=%f' % (yhat, obs))

```

MODEL BUILDING

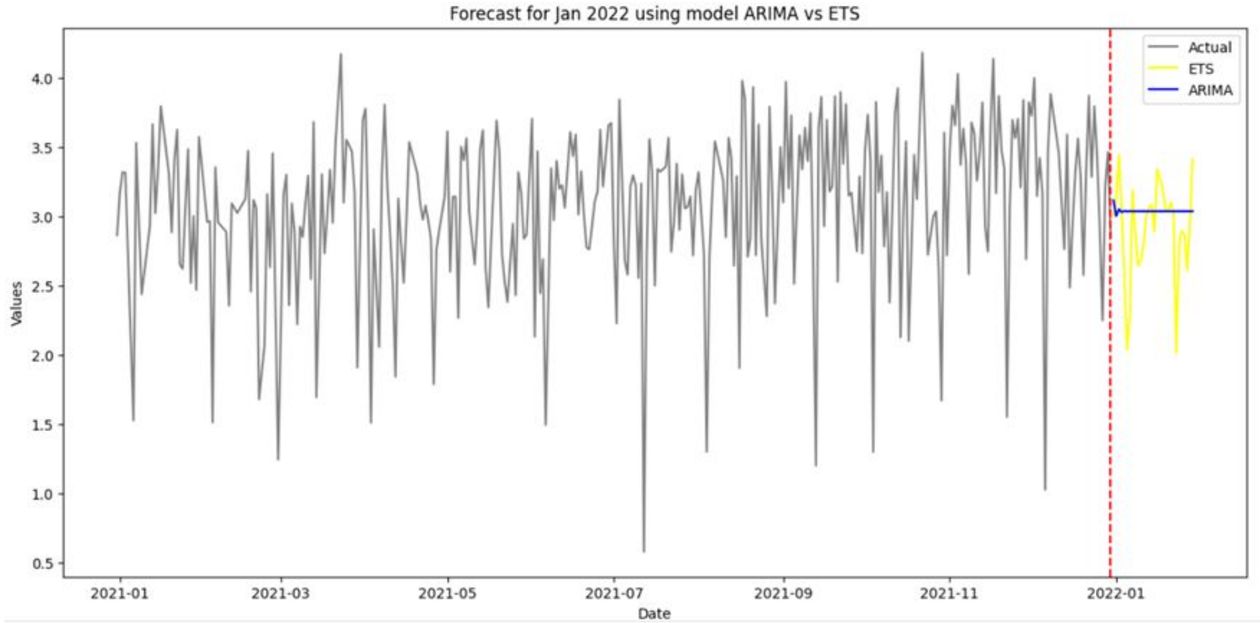
EVALUATION



	Model	RMSE
0	ARIMA	0.682820
1	ETS	0.569452

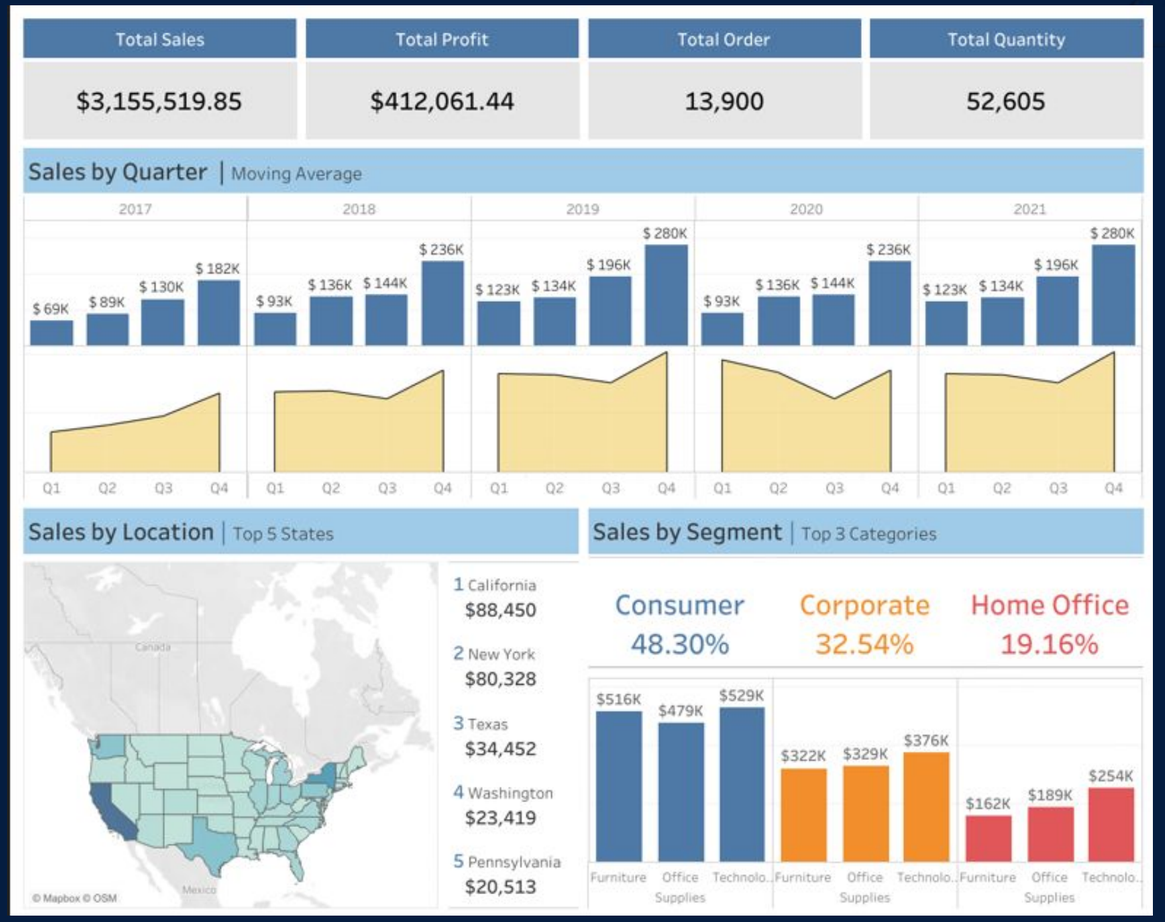
MODEL BUILDING

EVALUATION

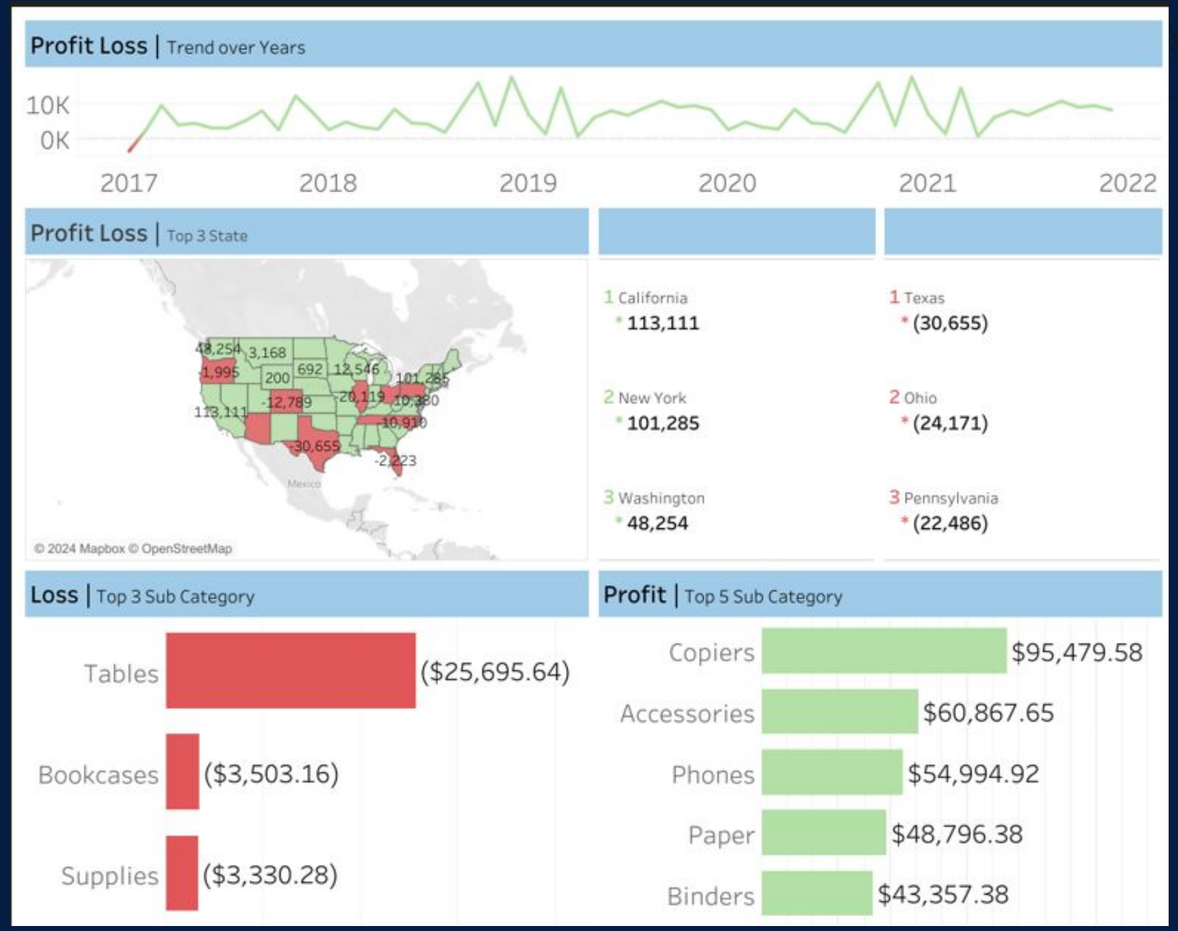


However, forecasting accuracy is critical. We can see that ARIMA forecast getting flat after a few days while ETS showing consistency through the end of month. Thus, ETS might be preferred over ARIMA based on above forecast for Jan 2022 and slightly lower RMSE value. We can conclude that the ETS model performs slightly better than the ARIMA model in terms of accuracy for this dataset

DASHBOARD


Total Sales
$3,155,519.85
Total Profit
$412,061.44
Total Order
13,900
Total Quantity
52,605
Sales by Quarter | Moving Average
2017
2018
2019
2020
2021
$ 69K
$ 89K
$ 130K
$ 182K
$ 93K
$ 136K
$ 144K
$ 236K
$ 123K
$ 134K
$ 196K
$ 280K
$ 93K
$ 136K
$ 144K
$ 236K
$ 123K
$ 134K
$ 196K
$ 280K
Q1 Q2 Q3 Q4
Sales by Location | Top 5 States
1 California
$88,450
2 New York
$80,328
3 Texas
$34,452
4 Washington
$23,419
5 Pennsylvania
$20,513
Canada
Mexico
© Mapbox © OSM
Sales by Segment | Top 3 Categories
Consumer
48.30%
Corporate
32.54%
Home Office
19.16%
$516K
$479K
$529K
$322K
$329K
$376K
$162K
$189K
$254K
Furniture
Office Supplies
Technolo..

DASHBOARD


Profit Loss | Trend over Years
10K
0K
2017
2018
2019
2020
2021
2022
Profit Loss | Top 3 State
48,254
3,168
1,995
200
692
12,546
101,285
20,119
10,380
12,789
113,111
10,910
30,655
2,223
Mexico
© 2024 Mapbox © OpenStreetMap
1 California
* 113,111
2 New York
* 101,285
3 Washington
* 48,254
1 Texas
* (30,655)
2 Ohio
* (24,171)
3 Pennsylvania
* (22,486)
Loss | Top 3 Sub Category
Tables
($25,695.64)
Bookcases
($3,503.16)
Supplies
($3,330.28)
Profit | Top 5 Sub Category
Copiers
$95,479.58
Accessories
$60,867.65
Phones
$54,994.92
Paper
$48,796.38
Binders
$43,357.38

CAPSTONE PROJECT

Thank You!