Time-Series Analysis & Profit Forecasting

Contents

- Dataset
- Objective
- Data Infrastructure
- Setting up environment
- Exploratory Data Analysis
- ARIMA/SARIMA statistical models
- Prophet Time-Series Forecasting
- Results
- Conclusion & Next Steps

Data

Preliminary Data Exploration

Below we can get an overview of the dataset and look at additional value we might be able to extract from the data.

[2]:
```
df= pd.read_csv('US Superstore data.csv')
df.head(6)
# Read in the dataset as 'df' and view top 4 rows
```

[2]:

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	...	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
0	1	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600	2	0.00	41.9136
1	2	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs,...	731.9400	3	0.00	219.5820
2	3	CA-2016-138688	2016-06-12	2016-06-16	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	...	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters b...	14.6200	2	0.00	6.8714
3	4	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	33311	South	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.0310
4	5	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.3680	2	0.20	2.5164
5	6	CA-2014-115812	2014-06-09	2014-06-14	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	...	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon Expressions Wood and Plastic Desk Access...	48.8600	7	0.00	14.1694

6 rows × 21 columns

Product Categories

Office Supplies	Furniture	Technology

Data Dictionary

'Row ID' - This is nothing but Serial No.
'Order ID' - ID created when a product order is placed.
'Order Date' - Date on which a customer places his/her order.
'Ship Date' - Date on which the order is shipped.
'Ship Mode' - Mode of shipment of each order.
'Customer ID' - ID assigned to each customer who places an order.
'Customer Name' - Name of Customer.
'Segment' - Section from where the order is placed.
'Country' - Country details of this data set. We are looking only for US store data.
'City' - Cities of US are listed here.
'State' - States of US are listed here.
'Postal Code' - pin code
'Region' - grouped into region wise
'Product ID' - Product ID of each product
'Category' - Category to which each product belongs to.
'Sub-Category' - Sub-Category of each Category
'Product Name' - Name of products.
'Sales' - Selling Price of each product.
'Quantity' - number of quantity available for a particular product.
'Discount' - Discount available on each product.
'Profit' - Profit gained on each product.

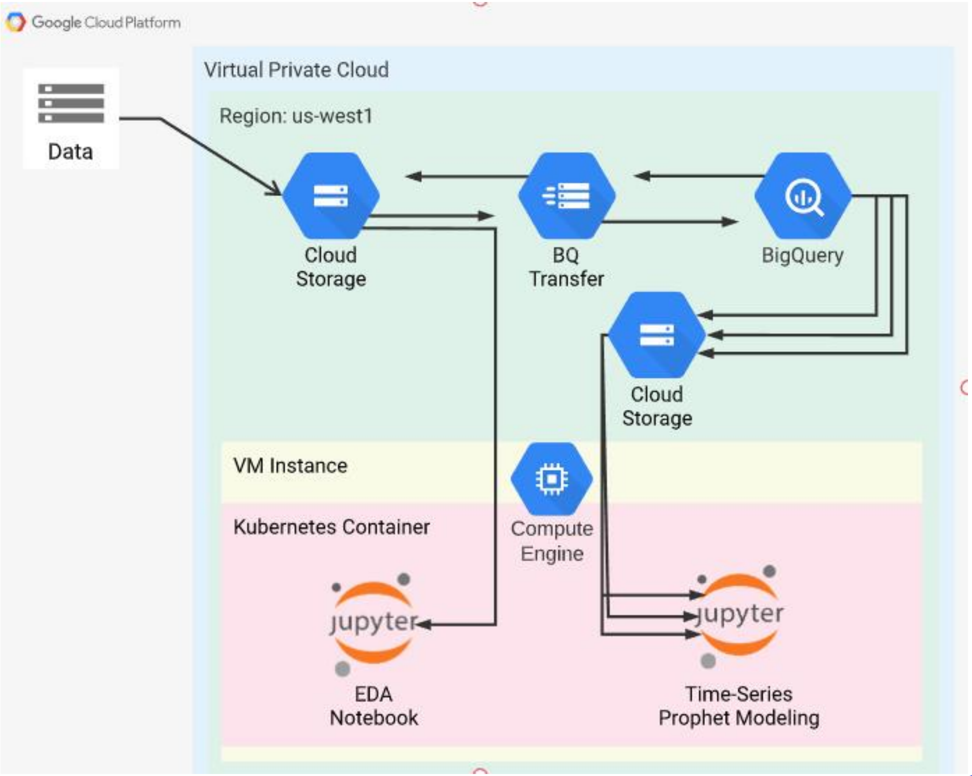
Data Source

Objective

"Time series forecasting is the use of a model to predict future values based on previously observed values."

The objective of this project is to determine the 'health' of all 3 product categories in this dataset. We want to understand and capture trends & seasonality, but also predict profits for each category for the next couple years. While doing so, I will explore some of the best models and statistical methods to work with and make predictions with time-series data.

Data Infrastructure



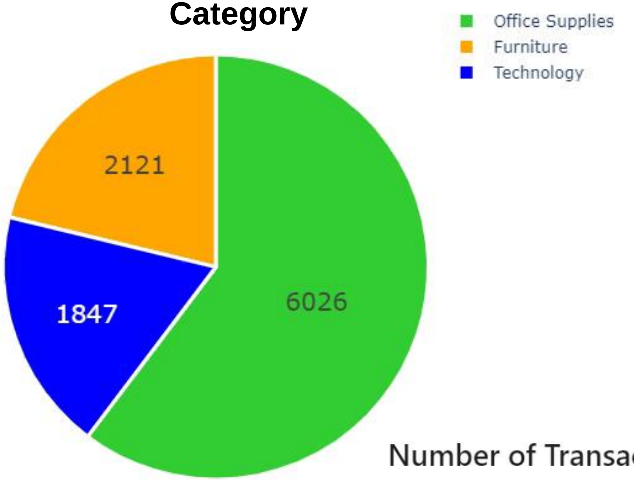
Setting up Environment

jupyter

Google Cloud Platform

Exploratory Data Analysis

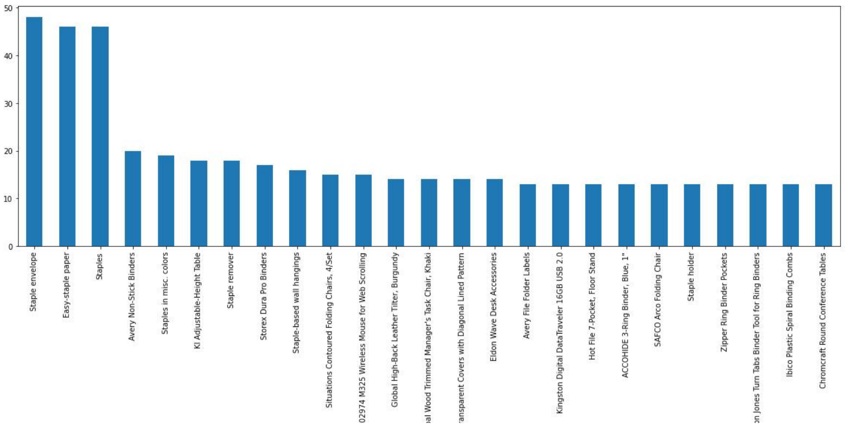
Transactions per Category



Products with the Most Transactions (Top 25)

```
prod_count=full_df[['Order ID','Customer ID','Product ID', 'Product Name', 'Sub-Category', 'Category', 'City','Quantity', 'Sales', 'Profit']]
product_sales_num=prod_count['Product Name'].value_counts()[:25]
product_sales_num.plot(kind='bar', figsize=(20,6))
# Set new df as 'prod_count', containing columns 'Product_Name', 'Sub-Category', 'Category', 'City','Quantity', 'Sales' & 'Profit'
# Check top 25 best selling products and assign it to 'product_sales_num'.
```

<AxesSubplot:>



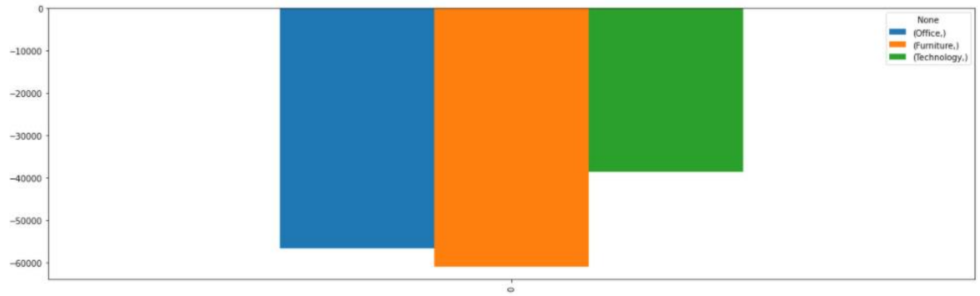
Number of Transactions per 'Sub-Category'

```
sub_cat=df['Sub-Category'].value_counts()
sub_cat=pd.DataFrame(sub_cat)
sub_cat=pd.DataFrame.transpose(sub_cat)
sub_cat

# Number of transactions per sub-categories
```

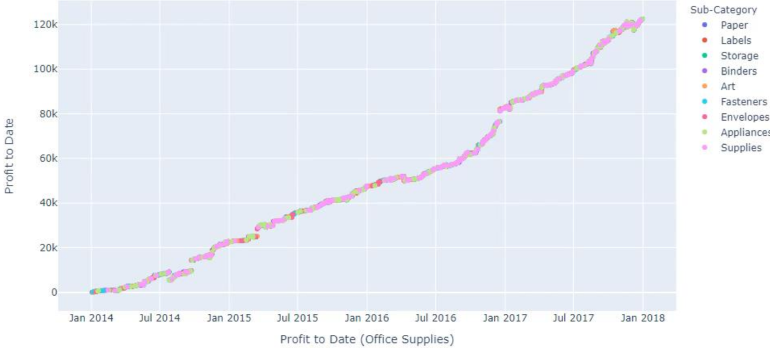
	Binders	Paper	Furnishings	Phones	Storage	Art	Accessories	Chairs	Appliances	Labels	Tables	Envelopes	Bookcases	Fasteners	Supplies	Machines	Copiers
Sub-Category	1523	1370	957	889	846	796	775	617	466	364	319	254	228	217	190	115	68

Exploring Profits


```
Total Profit Loss: -156131.2857
Total Office Supplies Profit Loss: -56615.258499999974
Total Furniture Profit Loss: -60936.109000000026
Total Technology Profit Loss: -38579.91820000001

<AxesSubplot:>
```



Plotting Profits Over Time



ARIMA/SARIMA Statistical Models

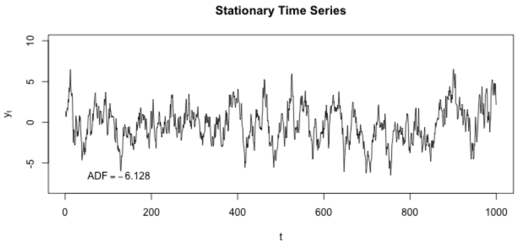
ARIMA: Autoregressive Integrated Moving Average
SARIMA: Seasonal Autoregressive Integrated Moving Average

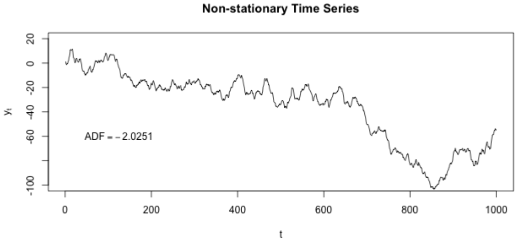
Autoregressive Model: Representation of a random process over a linear time scale and the output is dependant on previous values.

Integrated: Represents that the data is not stationary.

Moving Average: Calculating data points by taking average of previous forecasting errors.

Seasonality: A trend in data over some time period, typically 1 year.





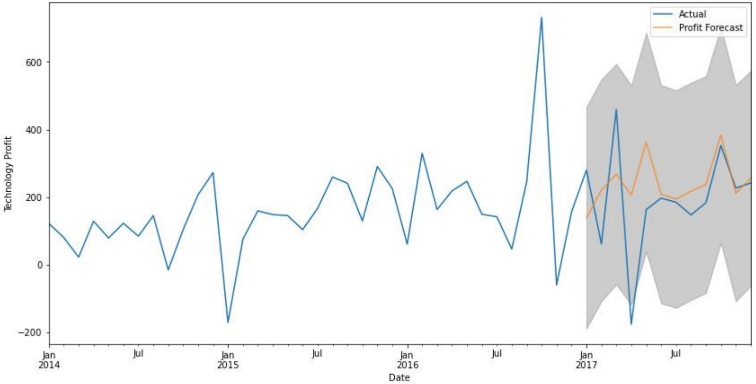
Resource

Implementing Models

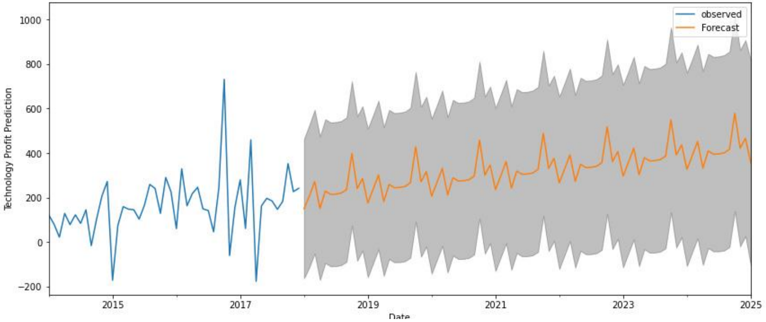
$$\underbrace{(1-\phi_1 B)}_{\text{Non-seasonal AR(1)}}\ \underbrace{(1-\Phi_1 B^4)}_{\text{Seasonal AR(1)}}\ \underbrace{(1-B)}_{\text{Non-seasonal difference}}\ \underbrace{(1-B^4)}_{\text{Seasonal difference}}y_t = \underbrace{(1+\theta_1 B)}_{\text{Non-seasonal MA(1)}}\ \underbrace{(1+\Theta_1 B^4)}_{\text{Seasonal MA(1)}}e_t.$$

SARIMA: (p,d,q) x (P,D,Q)s

```
pred_tech = results_tech.get_prediction(start=pd.to_datetime('2017-01-01'), dynamic=False)
pred_ci_t = pred_tech.conf_int()
#print(pred_ci_t)
axt = z['2014':].plot(label='Actual')
pred_tech.predicted_mean.plot(ax=axt, label='Profit Forecast', alpha=.7, figsize=(14, 7))
axt.fill_between(pred_ci_t.index,
                pred_ci_t.iloc[:, 0],
                pred_ci_t.iloc[:, 1], color='k', alpha=.2)
axt.set_xlabel('Date')
axt.set_ylabel('Technology Profit')
plt.legend()
plt.show()
z_predicted = pred_tech.predicted_mean
z_true = z['2017-01-01':]
mse = ((z_predicted - z_true)**2).mean()
print('Mean Square Error is:', round(mse, 4))
print('Root Mean Square Error is:', np.sqrt(mse))
# Technology
pred_uc_t = results_tech.get_forecast(steps=85)
pred_ci_t = pred_uc_t.conf_int()
ax = z.plot(label='observed', figsize=(14, 6))
pred_uc_t.predicted_mean.plot(ax=ax, label='Forecast')
ax.fill_between(pred_ci_t.index,
                pred_ci_t.iloc[:, 0],
                pred_ci_t.iloc[:, 1], color='k', alpha=.25)
ax.set_xlabel('Date')
ax.set_ylabel('Technology Profit Prediction')
plt.legend()
plt.show()
# Forecasting future Technology Profits.
# Technology Seasonality Pattern and Profit Prediction
```




```
Mean Square Error is: 23086.7302
Root Mean Square Error is: 151.94318078431314
```



Predicting Profit with Prophet

PROPHET

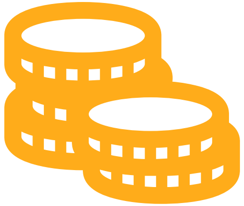
Prophet was made open source by Facebook, for the purposes of time-series forecasting and the model looks at non-linear trends in seasonality. It focuses on 3 main components Trend, Seasonality and Holidays.

Data Input format:

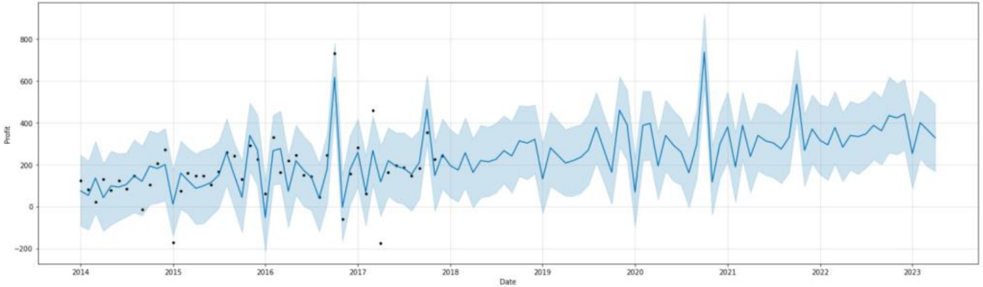
	ds	y
0	2014-01-01	46.408859
1	2014-02-01	10.358294
2	2014-03-01	57.746059
3	2014-04-01	41.675358
4	2014-05-01	79.418382
5	2014-06-01	137.792391
6	2014-07-01	-124.100860
7	2014-08-01	131.790510
8	2014-09-01	248.131119
9	2014-10-01	41.394096

Data Output:

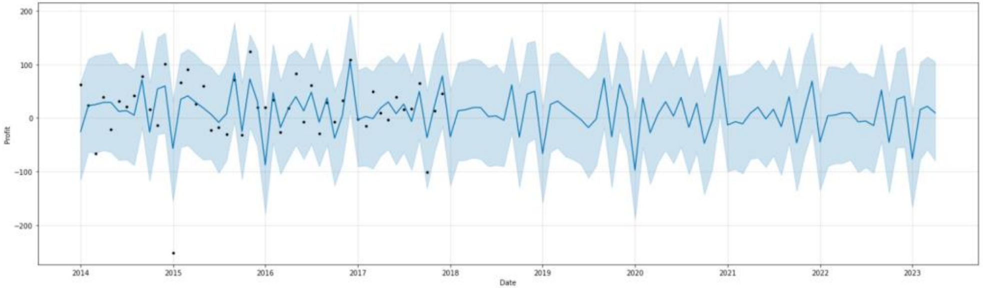
	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	additive_terms	additive_terms_lower	additive_terms_upper	yearly	yearly_lower	yearly_upper	multiplicative_terms	multiplicative_terms_lower	multiplicative_terms_upper	yhat
0	2014-01-01	80.622789	-50.216617	175.957473	80.622789	80.622789	-21.906807	-21.906807	-21.906807	-21.906807	-21.906807	-21.906807	0.0	0.0	0.0	58.715982
1	2014-02-01	82.062920	-90.965055	141.656070	82.062920	82.062920	-51.637080	-51.637080	-51.637080	-51.637080	-51.637080	-51.637080	0.0	0.0	0.0	30.425840
2	2014-03-01	83.363683	-14.802964	204.362627	83.363683	83.363683	18.547316	18.547316	18.547316	18.547316	18.547316	18.547316	0.0	0.0	0.0	101.911000
3	2014-04-01	84.803814	-88.994826	124.587709	84.803814	84.803814	-65.395075	-65.395075	-65.395075	-65.395075	-65.395075	-65.395075	0.0	0.0	0.0	19.408739
4	2014-05-01	86.197489	-60.667608	156.973677	86.197489	86.197489	-39.658131	-39.658131	-39.658131	-39.658131	-39.658131	-39.658131	0.0	0.0	0.0	46.539358



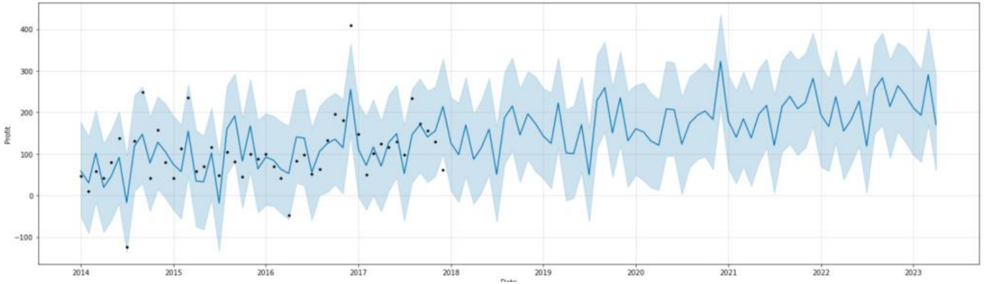
Technology Forecast



Furniture Forecast

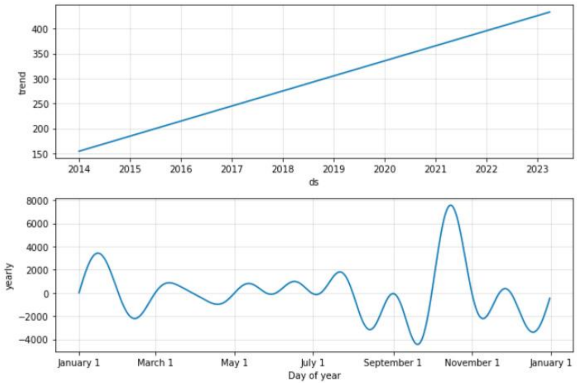


Office Supplies Forecast

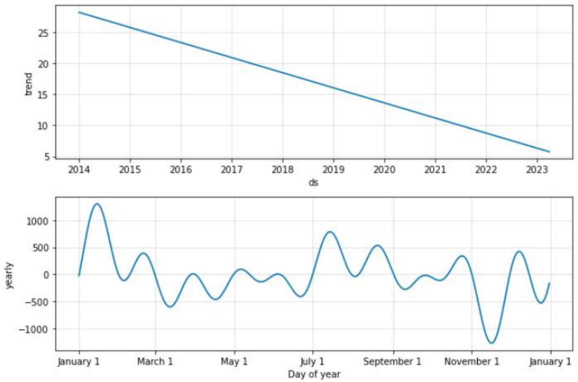


Trend & Seasonality

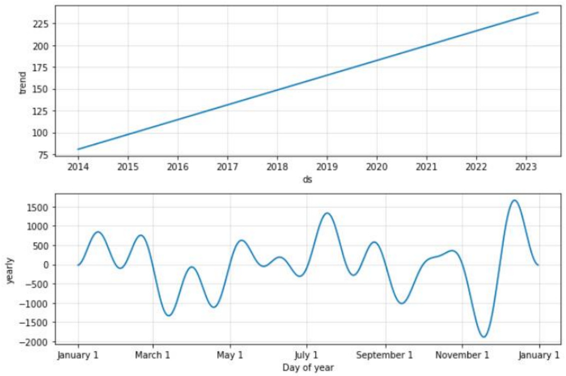
Technology Trend & Seasonality



Furniture Trend & Seasonality



Office Supplies Trend & Seasonality



Comparing Models

	Evaluation	Technology	Office Supplies	Furniture
ARIMA/SARIMA	MSE:	23086.73	4911.82	3145.44
	RMSE:	151.94	70.08	56.08
Prophet	MSE:	7209.14	3308.49	2108.26
	RMSE:	84.91	57.52	46.69

Conclusion & Next Steps

In conclusion we are able to get a pretty good understanding of the overall health of the 3 product categories and we are able to determine that technology and office supplies are trending upward and profits are expected to grow year over year. However when it comes to the furniture category, we can see that it is not trending upward in regards to profits. It was fairly clear from the ARIMA/SARIMA models but made even more apparent utilizing the Prophet model.

Next Steps:

- Automate GCP Pipeline
- Investigate Furniture category reasons for poor performance
- Prophet modeling for sub-categories