INTERNSHIP: INTERIM PROJECT REPORT

Dear Intern

Interim project report is an inherent component of your internship. We are enclosing a reference table of content for the interim project report.

The key objective of this report is for you to capture how far you have got in completing the internship work against milestones expected to be achieved within a specific duration and seek the mentor's feedback. Depending on the internship project and your progress (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections or leave some sections blank from the table of content mentioned below. You can also add additional sections. You can refer the project presentation to view the milestones related to your internship project. Please populate milestone# (1 / 2 / 3) and the milestone description in the interim project report based on the milestone for which you are submitting the interim project report.

You can refer the project presentation to view the milestones related to your internship project.

Internship Project Title	TCS iON RIO-125: Forecasting System - Project Demand of Products at a Retail Outlet Based on Historical Data.
Name of the Company	TCS iON
Name of the Industry Mentor	Sir Himalaya Ashish
Name of the Institute	Symbiosis University of Applies Sciences

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
2/03/21	18/03/21	6		Project Reference Material and webinar

Milestone #	2	Milestone:	Student should be able to choose appropriate forecasting model for the dataset and make predictions.

TABLE OF CONTENT

INTERNSHIP PROGRAM 2021

ON

TCS iON RIO-125: Forecasting System - Project Demand of Products at a Retail Outlet Based on Historical Data.

TCS iON

6th March 2021

ACKNOWLEGDEMENT

The internship opportunity that I had with TCS iON was a great change for learning and understanding the intricacies of the subject of; and also, for personal as well as professional development. I am very obliged for having a chance to interact with so many professionals who guided me throughout the internship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training & Development Team of TCS iON who gave me an opportunity to carry out my internship at their esteemed organization. Also, I express my thanks to the team for making me understand the details of the Data Science and Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for spearing his valuable time in spite of his busy schedule.

I would also like to thank the team of TCS iON and my colleagues who made the working environment productive and very conducive.

OBJECTIVE:

The objective of this project is to build a forecasting system to predict demand of a product based on historical data.

TCS iON

INTRODUCTION

What Is Demand Forecasting in Retail?
A Guide for Growing Businesses.

Demand forecasting is a key component to every growing retail business. Without proper demand forecasting processes in place, it can be nearly impossible to have the right amount of stock on hand at any given time. Too much merchandise in the warehouse means more capital tied up in inventory, and not enough could lead to out-of-stocks — and push customers to seek solutions from your competitors.
So, what is demand forecasting? And how is demand forecasting done in retail? Below, we'll explain demand forecasting and how you can use it to support your retail business' sustainable growth.

Table of contents

What is demand forecasting?

Demand forecasting in retail is the act of using data and insights to predict how much of a specific product or service customers will want to purchase during a defined time period. This method of predictive analytics helps retailers understand how much stock to have on hand at a given time.

WHAT IS DEMAND FORECASTING IN ECONOMICS?

Demand forecasting in economics is a bit different than how a retailer might use demand forecasting in business. So what do we mean by demand forecasting in economics, and how does that differ from retail?
In economics, analysts look at demand in the market as a whole, often for a particular industry or product category. In retail, you'll look at the demand

for YOUR products specifically. Demand forecasting in economics can (and should) inform forecasting in retail.

What is demand forecasting in marketing?

Demand forecasting in marketing is another component for retailers to consider. Get your marketing and operations teams on the same page so that

they can share calendars, priorities and initiatives and be proactive in planning. Retail ops can't provide inventory analytics for extra demand from a marketing campaign if they don't know about it in the first place.

Why demand forecasting is important?

When explaining why demand forecasting is important, the answer spans across several areas of a retail business. One Retail Systems Research report found that nearly three-quarters of "winning" retailers rate demand forecasting technologies as "very important" to their business and their success.

How does demand forecasting contribute to growing businesses?

It mostly comes down to two things: becoming more cost-efficient and improving the customer experience.

How demand forecasting makes your business more cost-efficient

Almost every retail business is always looking for ways to cut costs. It's one of the easiest ways to maximize your profits. When you implement a proper demand forecasting process to your business, you're cutting costs in a few ways.

Firstly, you're reducing the amount of capital you have tied up in unneeded inventory. And the less stock on hand you have, the lower your holding costs.

Secondly, you're making sure you capitalize on every sale opportunity by not disappointing customers with out-of-stocks.

Those are the two most straightforward ways, but you can also use demand forecasting to operate a lean and agile business, only investing money in more stock when you need to. When you've forecasted demand, you can easily check in before the period's over to see if you're on target to hit your predicted sales. If you're looking shy of your goal, you can.

amp up marketing and advertising. If it looks like you've underestimated, you could reorder or prep yourself to cross-promote a related product.

How demand forecasting enhances the customer experience

Another quick way to improve profits? Improve the customer experience. Rather than raising prices, focusing on the end user of the product can lead to customer loyalty and referrals.

Let's go back to the most obvious: avoiding out-of-stocks that disappoint customers and lead them to your competitors. This is one of the most impactful ways to please customers.

Beyond simply having enough product to meet demand, you can also use forecasting to inform staffing decisions. While this is relevant to businesses needing e commerce management, it especially pertains to brick-and-mortar retailers. Customers who come to your store want to speak to an associate. And if no one's there to help them, this can make a poor impression on shoppers. Even online sellers need to prep staff accordingly, especially during busy selling periods, so as not to delay shipping and fulfillment.

Uses of demand forecasting

As mentioned earlier, demand forecasting impacts many areas of your retail business. Here are just a few use cases of demand forecasting for rapidly growing businesses needing multichannel management:

- Prepare accurate budgets and financial planning
- Make informed purchasing decisions
- Implement purchase order automations to avoid stock issues
- Gain a thorough, comprehensive understanding of your business
- Anticipate staffing needs
- Grow sustainably
- Measure progress towards business and sales objectives
- Streamline production process
- Plan advertising and marketing campaigns and budgets
- Enhance the customer experience (avoid out-of-stocks, backorders, late shipments, etc.)
- Resourcing and project management

TCS iON

How is demand forecasting done, accurately?

Rather than asking "how is demand forecasting done?", retailers should ask "how is demand forecasting done *most accurately?*" There are many flaws to every approach to estimating demand and forecasting. Even though we can't predict the future perfectly, using established methods can help you be more successful in your forecasting practices.

Demand forecasting is done most accurately when a business considers both internal and external data. Internal metrics may include historical sales numbers, ad spend, and website or foot traffic. Externally speaking, you're looking at factors like industry or consumer trends, the weather, and even your competitors.

To best explain demand forecasting, it's helpful to look at the different methods. Some of the most common demand forecasting techniques include:

- Qualitative forecasting
- Time series analysis
- Causal model

Qualitative forecasting

This type of forecasting is when a business anticipates demand based on qualitative data. Qualitative data sources could include industry experts and/or consultants, employees, focus groups, and competitive analysis, to name a few. Often, this data is subjective and based on intuition rather than hard numbers or facts.

- Market research
- Delphi Method
- Expert opinion
- Focus groups
- Historical analogy
- Panel consensus
- Surveys

Recommended for: businesses that have limited historical data; new product launches (especially if there's no other product like it on the

market); instances where the previous period is believed to differ drastically from the planned period (for example, the Tickle Me Elmo frenzy during the 1996 holiday season)

Time series analysis

The time series analysis is a more quantitative approach to demand and forecasting. Rather than expert opinions and "soft" data inputs, a time series analysis uses exact numbers as the basis for forecasting demand. It's a more mathematical approach to forecasting which uses numerical inputs and trends. Other quantitative forecasting methods include:

- The indicator approach
- Econometric modeling
- Trend analysis

- Seasonal adjustment
- Decomposition
- Graphical methods
- Life cycle modeling

Recommended for: retailers that have plenty of past sales data (especially if this data reveals year-over-year trends); seasonal items; seasonal selling periods; identifying cyclical sales trends

Causal model

The causal model accounts for demand forecasting factors that may change predicted demand. Demand forecasting factors are both controllable and uncontrollable:

Controllable demand factors	Uncontrollable demand factors
Marketing, sales and promotions	Weather
Price	Politics
Visual merchandising	Trends
Location	Competitors
	Economic and socioeconomic conditions
	Seasonality

Because the causal method of forecasting accounts for so many variables, it's also a more complex approach. Some of the factors, like the weather, can't be predicted as accurately as you might like. This includes a part guesswork, part data-driven approach to forecasting — and a lot of trust in your intuition.
Recommended for: data-driven retailers with lots of metrics; forecasting by specific product, category or SKU; retailers in volatile markets; multi-channel businesses with a diverse customer base; forecasting in association with marketing/advertising campaigns and promotions

Demand forecasting tips

Demand forecasting is half art, half science. The best approach is to account for qualitative and quantitative data, internal and external variables, and controllable and uncontrollable factors. Many assumptions must be made, as well as "guesstimations" based off your experiences.

That being said, there are a few tips for demand forecasting that you can apply to ensure you're doing it properly:

Establish a baseline: This should be the first task on your list, aside from establishing a goal or hypothesis that you'll want to achieve or answer with your forecast. Without having a baseline of data, you're solely going off of third-party information.
Preserve your data: Because using your own data is so valuable in demand forecasting, you'll also need to ensure the data is clean and accurate. Centralize your inventory information so that everything is synced and in a single location, and you'll mitigate discrepancies.
Invest in the right tools: Without the right tools, demand forecasting can be a tedious, manual process. Find the right inventory management software that integrates with your accounting, point-of-sale and other tools for the most comprehensive look at your business.

It's not always clear what to look for in an inventory system, so we created a guide to help.

TCS iON

How to calculate demand forecasting accuracy
It'd be remiss to explain demand forecasting without also describing how to calculate demand forecasting accuracy. After all, demand forecasting can be done by almost anyone — but it's not always done accurately. And if your forecast is inaccurate, then you risk making majorly impactful business decisions based off the wrong information.
To calculate demand forecasting accuracy, many retailers look at the Mean Absolute Deviation (MAD) and Mean Absolute Percent Error (MAPE).
MAD =
MAD is the average difference between the actual demand and forecasted demand. To calculate MAD, you'll subtract the forecasted demand from the actual demand. You can then average this number over several time periods to find out your overall MAD.

$$\frac{1}{n}\sum|Actual - Forecast|$$

Month	Actual	Forecast	Absolute Error
1	112.3	124.7	12.4
2	108.4	103.7	4.7
3	148.9	116.6	32.3
4	117.4	78.5	38.9
MAD			**22.08**

MAPE =

MAPE measures the rate of accuracy of your forecast and is calculated by subtracting the forecasted demand from the actual demand, and then dividing that number by the actual demand. To get the percentage, multiply by 100. Again, you'll calculate this for multiple time periods and determine the average to find out your MAPE.

TCS iON

$$\left(\frac{1}{n}\sum\frac{|Actual - Forecast|}{|Actual|}\right)*100$$

Month	Actual	Forecast	Absolute Percent Error
1	112.3	124.7	11.0%
2	108.4	103.7	4.3%
3	148.9	116.6	21.7%
4	117.4	78.5	33.1%
MAPE			**17.6%**

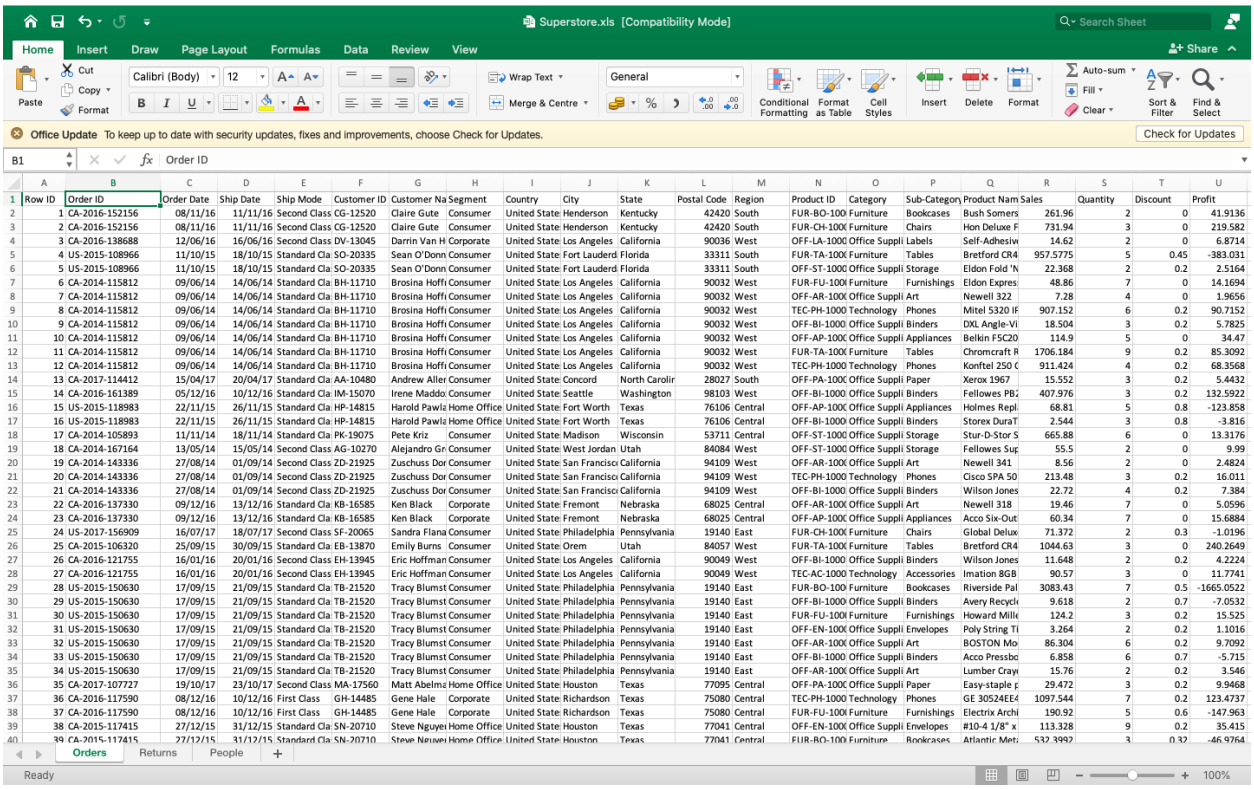
Business forecasting is essential for the survival for companies of all sizes. The building block used by forecasters is historical data or the past performance of the business to predict future results. Regression analysis is a statistical technique used to find relationships between independent and dependent variables. Regression analysis uses historical data and observation to predict future values.

TCS iON

APPROACH

1. Created a dataset, cleaned the dataset and also sanitized it.

- **Here is the screenshot of dataset, Superstore sales data.**



Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Na	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Categor	Product Nam	Sales	Quantity	Discount	Profit
1	CA-2016-152156	08/11/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-BO-100	Furniture	Bookcases	Bush Somers	261.96	2	0	41.9136
2	CA-2016-152156	08/11/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-CH-100	Furniture	Chairs	Hon Deluxe F	731.94	3	0	219.582
3	CA-2016-138688	12/06/16	16/06/16	Second Class	DV-13045	Darrin Van H	Corporate	United State	Los Angeles	California	90036	West	OFF-LA-1000	Office Suppli	Labels	Self-Adhesiv	14.62	2	0	6.8714
4	US-2015-108966	11/10/15	18/10/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	FUR-TA-100	Furniture	Tables	Bretford CR4	957.5775	5	0.45	-383.031
5	US-2015-108966	11/10/15	18/10/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	OFF-ST-1000	Office Suppli	Storage	Eldon Fold 'N	22.368	2	0.2	2.5164
6	CA-2014-115812	09/06/14	14/06/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	FUR-FU-100	Furniture	Furnishings	Eldon Expres	48.86	7	0	14.1694
7	CA-2014-115812	09/06/14	14/06/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	OFF-AR-100	Office Suppli	Art	Newell 322	7.28	4	0	1.9656
8	CA-2014-115812	09/06/14	14/06/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	TEC-PH-1000	Technology	Phones	Mitel 5320 IP	907.152	6	0.2	90.7152
9	CA-2014-115812	09/06/14	14/06/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	OFF-BI-1000	Office Suppli	Binders	DXL Angle-Vi	18.504	3	0.2	5.7825
10	CA-2014-115812	09/06/14	14/06/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	OFF-AP-100	Office Suppli	Appliances	Belkin F5C20	114.9	5	0	34.47
11	CA-2014-115812	09/06/14	14/06/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	FUR-TA-100	Furniture	Tables	Chromcraft R	1706.184	9	0.2	85.3092
12	CA-2014-115812	09/06/14	14/06/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	TEC-PH-1000	Technology	Phones	Konftel 250 C	911.424	4	0.2	68.3568
13	CA-2017-114412	15/04/17	20/04/17	Standard Cla	AA-10480	Andrew Aller	Consumer	United State	Concord	North Carolir	28027	South	OFF-PA-100	Office Suppli	Paper	Xerox 1967	15.552	3	0.2	5.4432
14	CA-2016-161389	05/12/16	10/12/16	Standard Cla	IM-15070	Irene Maddo	Consumer	United State	Seattle	Washington	98103	West	OFF-BI-1000	Office Suppli	Binders	Fellowes PB2	407.976	3	0.2	132.5922
15	US-2015-118983	22/11/15	26/11/15	Standard Cla	HP-14815	Harold Pawla	Home Office	United State	Fort Worth	Texas	76106	Central	OFF-AP-100	Office Suppli	Appliances	Holmes Repl	68.81	5	0.8	-123.858
16	US-2015-118983	22/11/15	26/11/15	Standard Cla	HP-14815	Harold Pawla	Home Office	United State	Fort Worth	Texas	76106	Central	OFF-BI-1000	Office Suppli	Binders	Storex DuraT	2.544	3	0.8	-3.816
17	CA-2014-105893	11/11/14	18/11/14	Standard Cla	PK-19075	Pete Kriz	Consumer	United State	Madison	Wisconsin	53711	Central	OFF-ST-1000	Office Suppli	Storage	Stur-D-Stor S	665.88	6	0	13.3176
18	CA-2014-167164	13/05/14	15/05/14	Second Class	AG-10270	Alejandro Gr	Consumer	United State	West Jordan	Utah	84084	West	OFF-ST-1000	Office Suppli	Storage	Fellowes Sup	55.5	2	0	9.99
19	CA-2014-143336	27/08/14	01/09/14	Second Class	ZD-21925	Zuschuss Dor	Consumer	United State	San Francisc	California	94109	West	OFF-AR-100	Office Suppli	Art	Newell 341	8.56	2	0	2.4824
20	CA-2014-143336	27/08/14	01/09/14	Second Class	ZD-21925	Zuschuss Dor	Consumer	United State	San Francisc	California	94109	West	TEC-PH-1000	Technology	Phones	Cisco SPA 50	213.48	3	0.2	16.011
21	CA-2014-143336	27/08/14	01/09/14	Second Class	ZD-21925	Zuschuss Dor	Consumer	United State	San Francisc	California	94109	West	OFF-BI-1000	Office Suppli	Binders	Wilson Jones	22.72	4	0.2	7.384
22	CA-2016-137330	09/12/16	13/12/16	Standard Cla	KB-16585	Ken Black	Corporate	United State	Fremont	Nebraska	68025	Central	OFF-AR-100	Office Suppli	Art	Newell 318	19.46	7	0	5.0596
23	CA-2016-137330	09/12/16	13/12/16	Standard Cla	KB-16585	Ken Black	Corporate	United State	Fremont	Nebraska	68025	Central	OFF-AP-100	Office Suppli	Appliances	Acco Six-Out	60.34	7	0	15.6884
24	US-2017-156909	16/07/17	18/07/17	Second Class	SF-20065	Sandra Flana	Consumer	United State	Philadelphia	Pennsylvania	19140	East	FUR-CH-100	Furniture	Chairs	Global Delux	71.372	2	0.3	-1.0196
25	CA-2015-106320	25/09/15	30/09/15	Standard Cla	EB-13870	Emily Burns	Consumer	United State	Orem	Utah	84057	West	FUR-TA-100	Furniture	Tables	Bretford CR4	1044.63	3	0	240.2649
26	CA-2016-121755	16/01/16	20/01/16	Second Class	EH-13945	Eric Hoffman	Consumer	United State	Los Angeles	California	90049	West	OFF-BI-1000	Office Suppli	Binders	Wilson Jones	11.648	2	0.2	4.2224
27	CA-2016-121755	16/01/16	20/01/16	Second Class	EH-13945	Eric Hoffman	Consumer	United State	Los Angeles	California	90049	West	TEC-AC-1000	Technology	Accessories	Imation 8GB	90.57	3	0	11.7741
28	US-2015-150630	17/09/15	21/09/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	FUR-BO-100	Furniture	Bookcases	Riverside Pal	3083.43	7	0.5	-1665.0522
29	US-2015-150630	17/09/15	21/09/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-BI-1000	Office Suppli	Binders	Avery Recycl	9.618	2	0.7	-7.0532
30	US-2015-150630	17/09/15	21/09/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	FUR-FU-100	Furniture	Furnishings	Howard Mill	124.2	3	0.2	15.525
31	US-2015-150630	17/09/15	21/09/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-EN-100	Office Suppli	Envelopes	Poly String Ti	3.264	2	0.2	1.1016
32	US-2015-150630	17/09/15	21/09/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-AR-100	Office Suppli	Art	BOSTON Mo	86.304	6	0.2	9.7092
33	US-2015-150630	17/09/15	21/09/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-BI-1000	Office Suppli	Binders	Acco Pressbo	6.858	6	0.7	-5.715
34	US-2015-150630	17/09/15	21/09/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-AR-100	Office Suppli	Art	Lumber Cray	15.76	2	0.2	3.546
35	CA-2017-107727	19/10/17	23/10/17	Second Class	MA-17560	Matt Abelma	Home Office	United State	Houston	Texas	77095	Central	OFF-PA-100	Office Suppli	Paper	Easy-staple p	29.472	3	0.2	9.9468
36	CA-2016-117590	08/12/16	10/12/16	First Class	GH-14485	Gene Hale	Corporate	United State	Richardson	Texas	75080	Central	TEC-PH-1000	Technology	Phones	GE 30524EE4	1097.544	7	0.2	123.4737
37	CA-2016-117590	08/12/16	10/12/16	First Class	GH-14485	Gene Hale	Corporate	United State	Richardson	Texas	75080	Central	FUR-FU-100	Furniture	Furnishings	Electrix Archi	190.92	5	0.6	-147.963
38	CA-2015-117415	27/12/15	31/12/15	Standard Cla	SN-20710	Steve Nguyer	Home Office	United State	Houston	Texas	77041	Central	OFF-EN-100	Office Suppli	Envelopes	#10-4 1/8" x	113.328	9	0.2	35.415
39	CA-2015-117415	27/12/15	31/12/15	Standard Cla	SN-20710	Steve Nguyer	Home Office	United State	Houston	Texas	77041	Central	FUR-BO-100	Furniture	Bookcases	Atlantic Met	532.3992	3	0.32	-46.9764

TCS iON

- **There are several categories in superstore sales data, we start from time series analysis and forecasting for furniture sales.**

jupyter forecasting sales Last Checkpoint: Last Thursday at 23:26 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

In [1]:
```
import warnings
import itertools
import numpy as np
import matplotlib.pyplot as plt
warnings.filterwarnings("ignore")
plt.style.use('fivethirtyeight')
import pandas as pd
import statsmodels.api as sm
import matplotlib
matplotlib.rcParams['axes.labelsize'] = 14
matplotlib.rcParams['xtick.labelsize'] = 12
matplotlib.rcParams['ytick.labelsize'] = 12
matplotlib.rcParams['text.color'] = 'k'
```

In [8]:
```
df = pd.read_excel("Superstore.xls")
furniture = df.loc[df['Category'] == 'Furniture']
df.head()
```

Out[8]:

	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	Produ Nam
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	Bus Somers Collectic Bookcas
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 Delux Fabr Upholstere Stackir Chairs,
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	Sel Adhesiv Addres Labels f Typewrite b
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	Bretfo CR45(Series Sli Rectangul Tab

- **We have good 4-year sales furniture sales data.**

In [5]:
```
furniture['Order Date'].min(), furniture['Order Date'].max()
```

Out[5]: (Timestamp('2014-01-06 00:00:00'), Timestamp('2017-12-30 00:00:00'))

TCS iON

- **This step includes removing columns we do not need, check missing values, aggregate sales by date and so on.**

```
In [6]: cols = ['Row ID', 'Order ID', 'Ship Date', 'Ship Mode', 'Customer ID', 'Customer Name', 'Segment', 'Country', 'City'
        furniture.drop(cols, axis=1, inplace=True)
        furniture = furniture.sort_values('Order Date')
        furniture.isnull().sum()

Out[6]: Order Date    0
        Sales         0
        dtype: int64
```

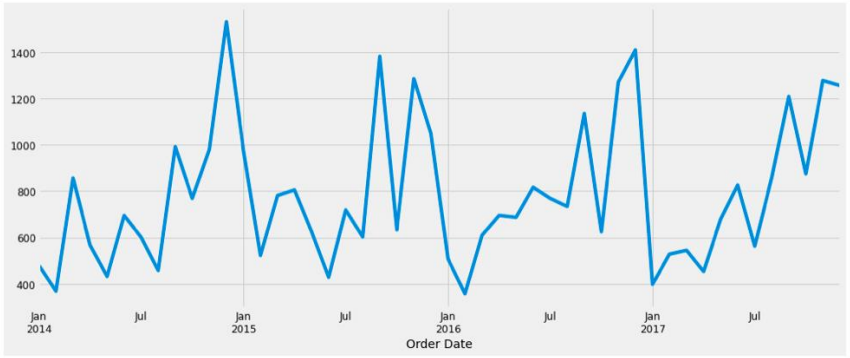
- **Indexing, our current datetime data can be tricky to work with, therefore, we will use the averages daily sales value for that month instead, and we are using the start of each month as the timestamp.**

```
In [6]: furniture = furniture.set_index('Order Date')
        furniture.index

Out[6]: DatetimeIndex(['2014-01-06', '2014-01-07', '2014-01-10', '2014-01-11',
                       '2014-01-13', '2014-01-14', '2014-01-16', '2014-01-19',
                       '2014-01-20', '2014-01-21',
                       ...
                       '2017-12-18', '2017-12-19', '2017-12-21', '2017-12-22',
                       '2017-12-23', '2017-12-24', '2017-12-25', '2017-12-28',
                       '2017-12-29', '2017-12-30'],
                      dtype='datetime64[ns]', name='Order Date', length=889, freq=None)

In [7]: y = furniture['Sales'].resample('MS').mean()

In [8]: y['2017':]

Out[8]: Order Date
        2017-01-01     397.602133
        2017-02-01     528.179800
        2017-03-01     544.672240
        2017-04-01     453.297905
        2017-05-01     678.302328
        2017-06-01     826.460291
        2017-07-01     562.524857
        2017-08-01     857.881889
        2017-09-01    1209.508583
        2017-10-01     875.362728
        2017-11-01    1277.817759
        2017-12-01    1256.298672
        Freq: MS, Name: Sales, dtype: float64
```

TCS iON

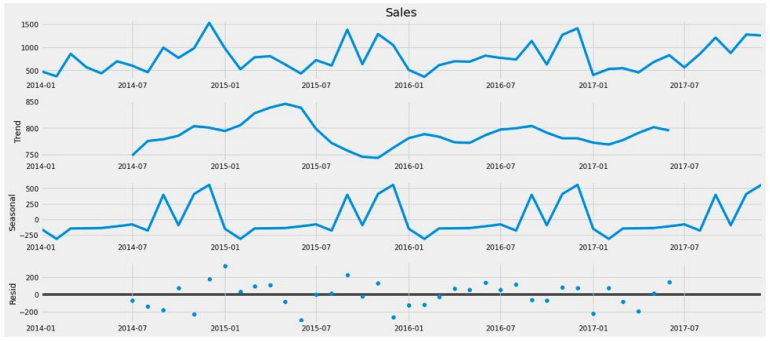
- **Visualizing furniture sales time series data**

```
In [9]: y.plot(figsize=(15, 6))
        plt.show()
```



- **Some distinguishable patterns appear when we plot the data. The time-series has seasonality pattern, such as sales are always low at the beginning of the year and high at the end of the year. There is always an upward trend within any single year with a couple of low months in the mid of the year. We can also visualize our data using a method called time-series decomposition that allows us to decompose our time series into three distinct components: trend, seasonality, and noise.**

```
In [10]: from pylab import rcParams
rcParams['figure.figsize'] = 18, 8
decomposition = sm.tsa.seasonal_decompose(y, model='additive')
fig = decomposition.plot()
plt.show()
```



ARIMA (Auto-Regressive Integrated Moving Average) Model

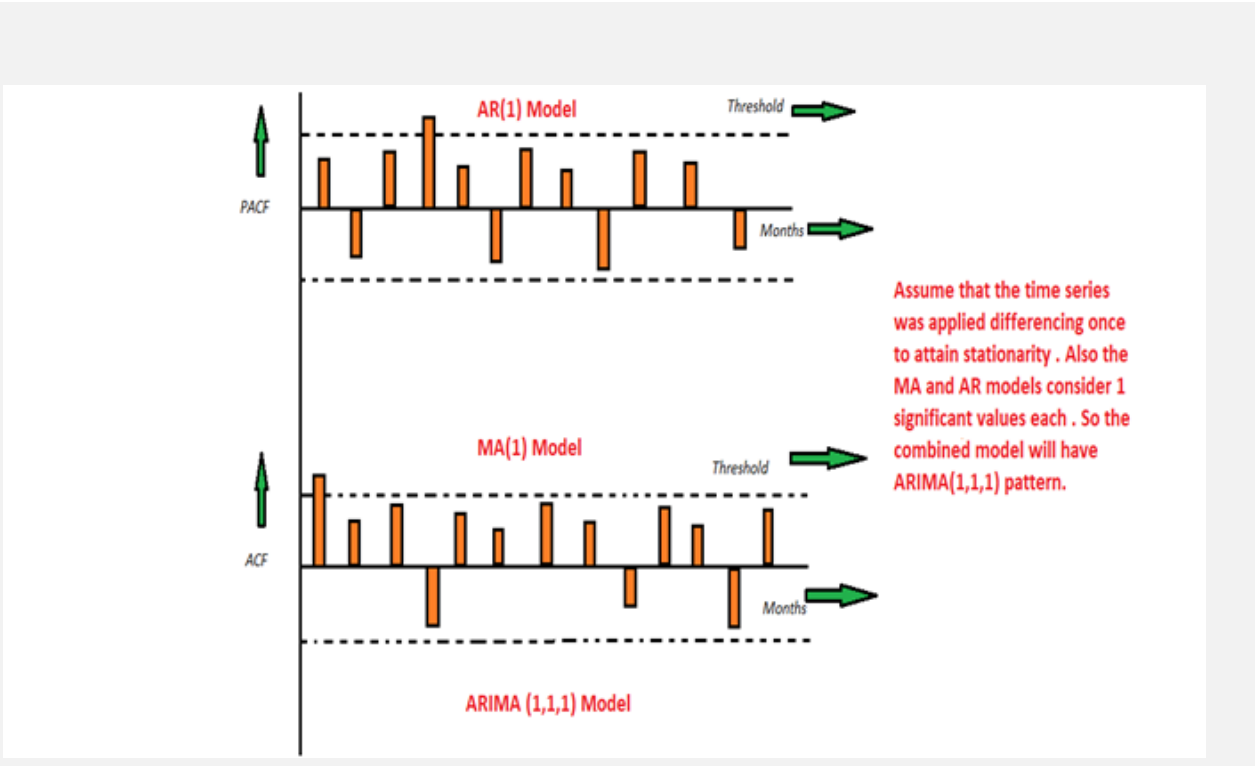


Image by Author

We know that in order to apply the various models we must in the beginning convert the series into Stationary Time Series. In order to achieve the same, we apply the differencing or Integrated method where we subtract the t-1 value from t values of time series. After applying the first differencing if we are still unable to get the Stationary time series then we again apply the second-order differencing.

The ARIMA model is quite similar to the ARMA model other than the fact that it includes one more factor known as Integrated(I) i.e. differencing which stands for I in the ARIMA model. So in short ARIMA model is a combination of a number of differences already applied on the model in order to make it stationary, the number of previous lags along with residuals errors in order to forecast future values.

Consider the above graphs where the MA and AR values are plotted with their respective significant values. Let's assume that we consider only 1 significant value from the AR model and likewise 1 significant value from the MA model. Also, the graph was initially non-stationary and we had to perform differencing operation once in order to convert into a stationary set. Hence the ARIMA model which will be obtained from the combined values of the other two models along with the Integral operator can be displayed as ARIMA(1,1,1).

TCS iON

Python Code

```
1  # ARIMA example
2  from statsmodels.tsa.arima.model import ARIMA
3  from random import random
4  # contrived dataset
5  data = [x + random() for x in range(1, 100)]
6  # fit model
7  model = ARIMA(data, order=(1, 1, 1))
8  model_fit = model.fit()
9  # make prediction
10 yhat = model_fit.predict(len(data), len(data), typ='levels')
11 print(yhat)
```

TIME SERIES FORECASTING WITH ARIMA

We are going to apply one of the most commonly used method for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average.

ARIMA models are denoted with the notation `ARIMA(p, d, q)`. These three parameters account for seasonality, trend, and noise in data:

Time series forecasting with ARIMA

In [11]:
```
#ARIMA models are denoted with the notation ARIMA(p, d, q). These three parameters account for seasonality, trend, a
p = d = q = range(0, 2)
pdq = list(itertools.product(p, d, q))
seasonal_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, d, q))]
print('Examples of parameter combinations for Seasonal ARIMA...')
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[1]))
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[2]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[3]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[4]))
```

```
Examples of parameter combinations for Seasonal ARIMA...
SARIMAX: (0, 0, 1) x (0, 0, 1, 12)
SARIMAX: (0, 0, 1) x (0, 1, 0, 12)
SARIMAX: (0, 1, 0) x (0, 1, 1, 12)
SARIMAX: (0, 1, 0) x (1, 0, 0, 12)
```

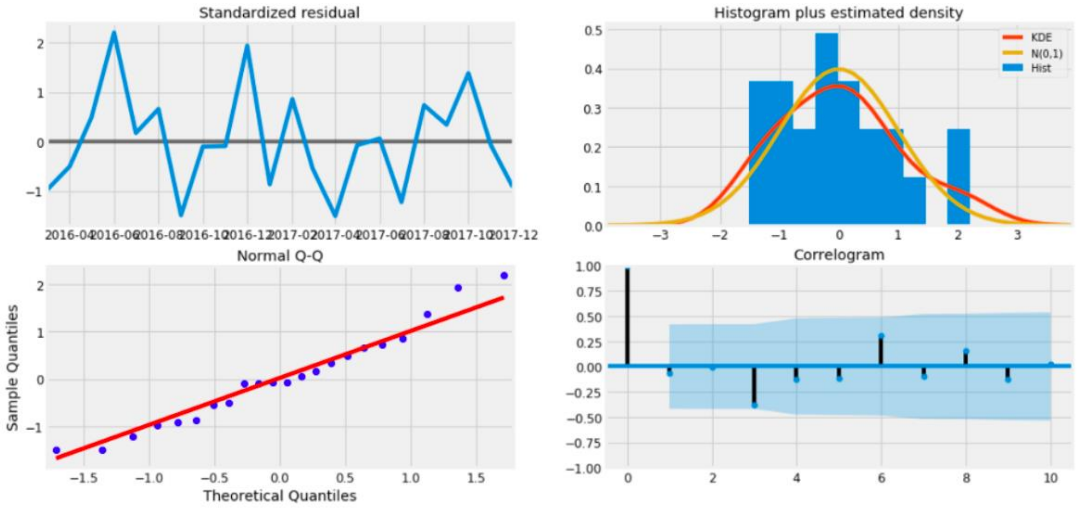
In [19]:
```
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod = sm.tsa.statespace.SARIMAX(y,
                                            order=param,
                                            seasonal_order=param_seasonal,
                                            enforce_stationarity=False,
                                            enforce_invertibility=False)

            results = mod.fit()

            print('ARIMA{}x{}12 - AIC:{}'.format(param, param_seasonal, results.aic))
        except:
            continue
```

```
ARIMA(0, 0, 0)x(0, 0, 0, 12)12 - AIC:769.0817523205915
ARIMA(0, 0, 0)x(0, 0, 1, 12)12 - AIC:1436.7096275959661
ARIMA(0, 0, 0)x(0, 1, 0, 12)12 - AIC:477.7170130920218
/opt/anaconda3/lib/python3.7/site-packages/statsmodels/base/model.py:568: ConvergenceWarning: Maximum Likelihood op
timization failed to converge. Check mle_retvals
  ConvergenceWarning)
ARIMA(0, 0, 0)x(0, 1, 1, 12)12 - AIC:302.2702899793748
ARIMA(0, 0, 0)x(1, 0, 0, 12)12 - AIC:497.23144334183365
/opt/anaconda3/lib/python3.7/site-packages/statsmodels/base/model.py:568: ConvergenceWarning: Maximum Likelihood op
timization failed to converge. Check mle_retvals
  ConvergenceWarning)
ARIMA(0, 0, 0)x(1, 0, 1, 12)12 - AIC:1131.1862176526117
ARIMA(0, 0, 0)x(1, 1, 0, 12)12 - AIC:318.0047199116341
ARIMA(0, 0, 0)x(1, 1, 1, 12)12 - AIC:304.2488280301923
ARIMA(0, 0, 1)x(0, 0, 0, 12)12 - AIC:720.9252270758111
```

FITTING THE ARIMA MODEL

```
In [20]: mod = sm.tsa.statespace.SARIMAX(y,
                                         order=(1, 1, 1),
                                         seasonal_order=(1, 1, 0, 12),
                                         enforce_stationarity=False,
                                         enforce_invertibility=False)
         results = mod.fit()
         print(results.summary().tables[1])
```

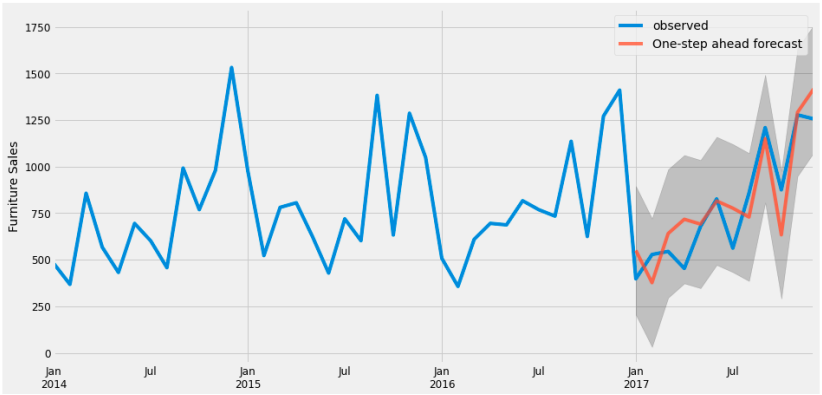
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0146	0.342	0.043	0.966	-0.655	0.684
ma.L1	-1.0000	0.360	-2.781	0.005	-1.705	-0.295
ar.S.L12	-0.0253	0.042	-0.609	0.543	-0.107	0.056
sigma2	2.958e+04	1.22e-05	2.43e+09	0.000	2.96e+04	2.96e+04



- ## VALIDATING FORECASTS

To help us understand the accuracy of our forecasts, we compare predicted sales to real sales of the time series, and we set forecasts to start at 2017–01–01 to the end of the data.

```
In [17]: pred=results.get_prediction(start=pd.to_datetime('2017-01-01'),dynamic=False)
         pred_ci=pred.conf_int()

         ax=y['2014':].plot(label='observed')
         pred.predicted_mean.plot(ax=ax,label='One-step ahead forecast', alpha=.7, figsize=(14,7))
         ax.fill_between(pred_ci.index,
                         pred_ci.iloc[:,0],
                         pred_ci.iloc[:,1],color='k',alpha=.2)

         ax.set_xlabel('Date')
         ax.set_ylabel('Furniture Sales')
         plt.legend()

         plt.show()
```



MEAN SQUARED ERROR

Now consider we are using SSE as our loss function. So if we have a dataset of say 100 points, our SSE is, say, 200. If we increased data points to 500, our SSE would increase as the squared errors will add up for 500 data points now. So let's say it becomes 800. If we increase the number of data points again, our SSE will further increase. Fair enough? Absolutely not!

N	L
100	200
500	800
1000	1200

The error should decrease as we increase our sample data as the distribution of our data becomes more and more narrower (referring to normal distribution). The more data we have, the less is the error. But in the case of SSE, the complete opposite is happening. Here, finally, comes in our warrior — Mean Squared Error. Its expression is:

$$\mathbf{L} = \frac{1}{N}[\Sigma(\hat{Y} - Y)^2]$$

We take the average or mean of SSE. So more the data, lesser will be the aggregated error, MSE.

N	L	L/N
100	200	2
500	800	1.6
1000	1200	1.2

Here as you can see, the error is decreasing as our algorithm is gaining more and more *experience*. The Mean Squared Error is used as a default metric for evaluation of the performance of most regression algorithms be it R, Python or even MATLAB.

ROOT MEAN SQUARED ERROR

The only issue with MSE is that the order of loss is more than that of the data. As my data is of order 1 and the loss function, MSE has an order of 2. So we cannot directly correlate data with the error. Hence, we take the root of the MSE — which is the Root Mean Squared Error:

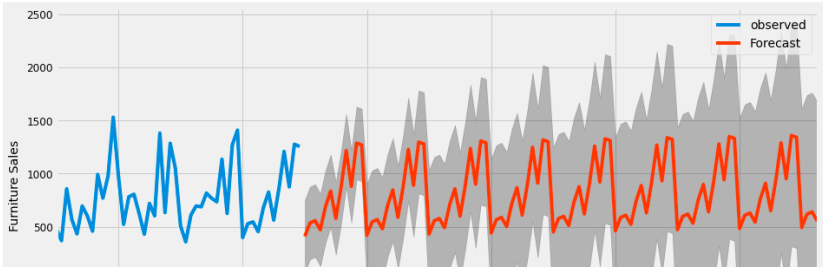
$$\mathbf{L} = \sqrt{\frac{1}{N}[\Sigma(\hat{Y} - Y)^2]}$$

Here, we are not changing the loss function and the solution is still the same. All we have done is reduced the order of the loss function by taking the root.

In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. The MSE is a measure of the quality of an estimator — it is always non-negative, and the smaller the MSE, the closer we are to finding the line of best fit.

Root Mean Square Error (RMSE) tells us that our model was able to forecast the average daily furniture sales in the test set within 151.64 of the real sales. Our furniture daily sales range from around 400 to over 1200. In my opinion, this is a pretty good model so far.

```
In [18]: y_forecasted = pred.predicted_mean
         y_truth = y['2017-01-01':]
         mse = ((y_forecasted - y_truth) ** 2).mean()
         print('The Mean Squared Error of our forecasts is {}'.format(round(mse, 2)))

         The Mean Squared Error of our forecasts is 22993.56

In [19]: print('The Root Mean Squared Error of our forecasts is {}'.format(round(np.sqrt(mse), 2)))

         The Root Mean Squared Error of our forecasts is 151.64

In [20]: pred_uc = results.get_forecast(steps=100)
         pred_ci = pred_uc.conf_int()
         ax = y.plot(label='observed', figsize=(14, 7))
         pred_uc.predicted_mean.plot(ax=ax, label='Forecast')
         ax.fill_between(pred_ci.index,
                         pred_ci.iloc[:, 0],
                         pred_ci.iloc[:, 1], color='k', alpha=.25)
         ax.set_xlabel('Date')
         ax.set_ylabel('Furniture Sales')
         plt.legend()
         plt.show()
```



Conclusion:

All these models give us an insight or at least close enough prediction about any particular time series. Also, it depends on the users that which model perfectly suffices their needs. If the chances of error rate are less in any one model compared to other models then it's preferred that we choose the one which gives us the closest estimation.

THANK YOU