CECS 551: Advanced Artificial Intelligence
Group ~ 2

Final Project Presentation

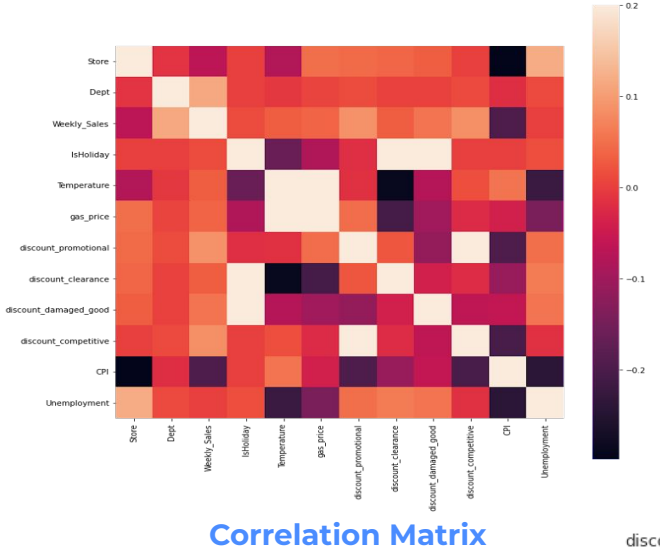
Datasets used:

- **CECS551 dataset 01:** The task is to predict the department-wide sales for each store.
- **CECS551 dataset 02**: The goal is to predict the *unit* sales of each product for the next 10 days from 10 different stores across various states.

Instructor:
Professor Mahshid Fardadi

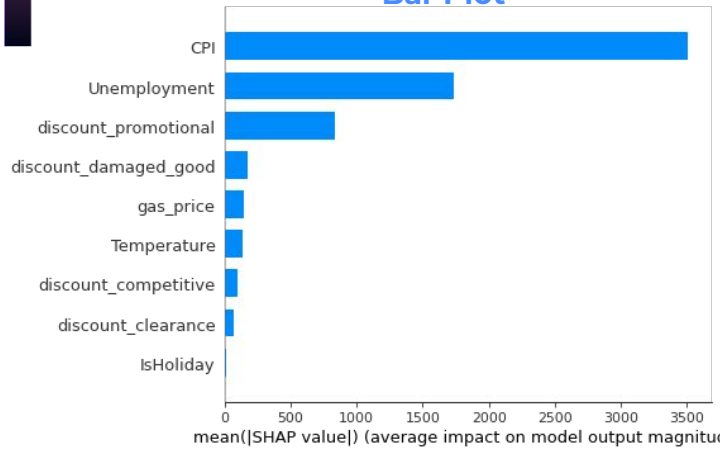
Group Members:
Pragya Sharma
Suraj Shah
Harsha
Taher Vora
Almee Christian

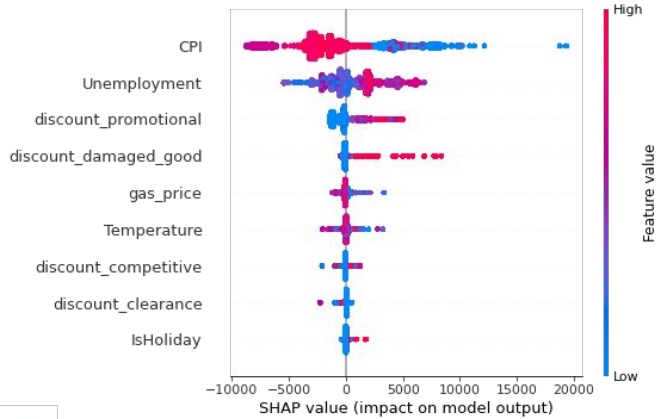
| Feature Importance



Correlation Matrix

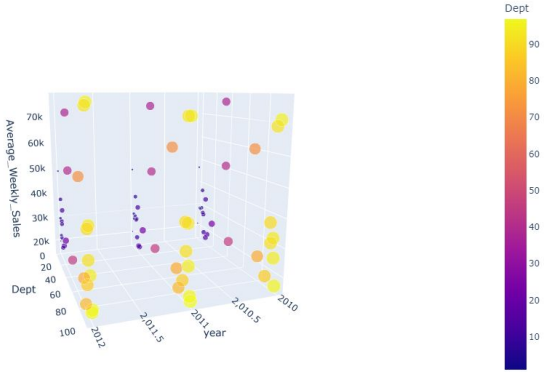
Bar Plot



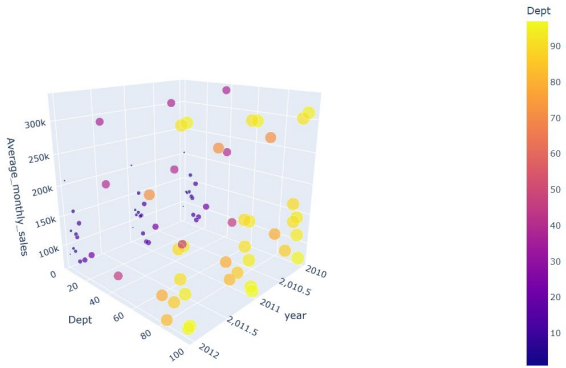


BeesWarm Plot

| Weekly and monthly sales patterns for top 35% of the department sales.



Average Weekly Sales



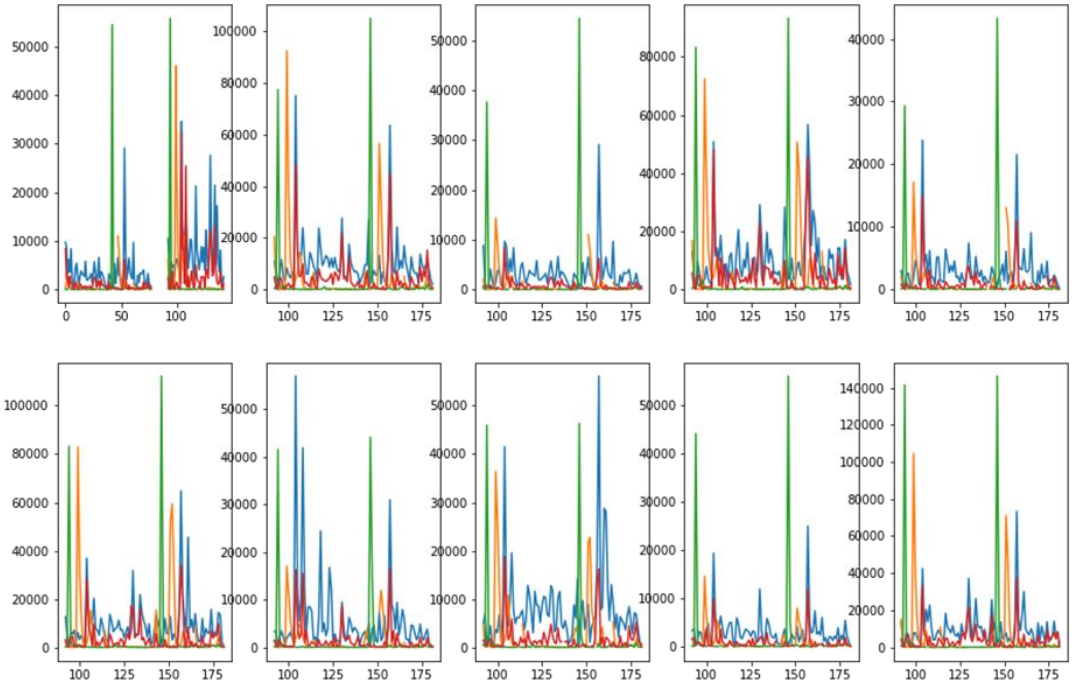
Average Monthly Sales

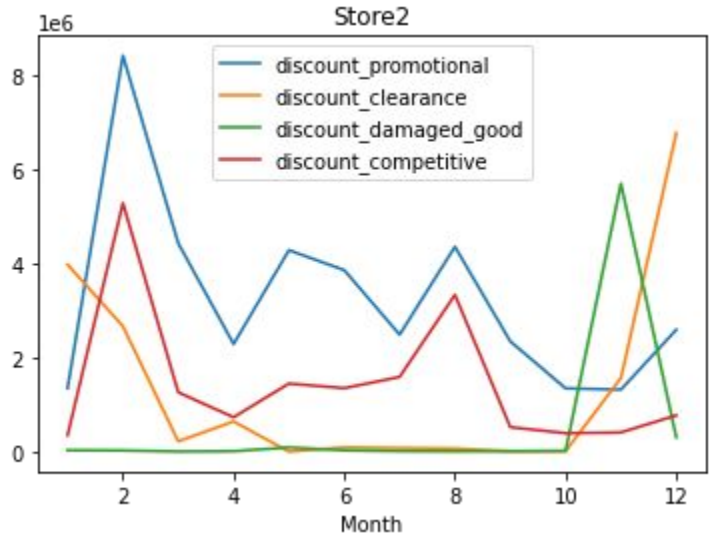
Conclusion:

Department 95 is the best department across all the 10 stores as it shows consistently increasing high average sales.

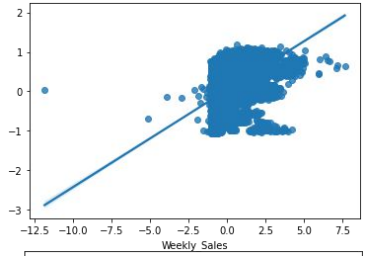
| Impact of Various discounts on overall Sales

- Plot to the right shows discount of 10 stores where we can see that discount_promotional plays most important role in sales.
- Plot at bottom shows one the best of 30% performing stores where again it implies that discount_promotional plays an important role in sales.





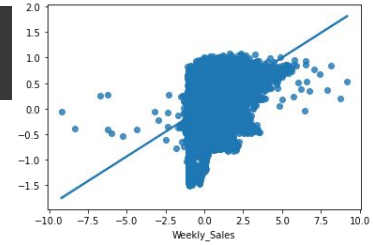
Stores 1 to 10



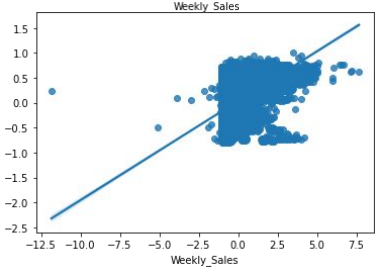

```
MSE: 0.7880411373245544
MAE: 0.6246598472526079
R2 : 0.2508821045622597
```

Stores 11 to 35

```
MSE: 0.7393404224768703
MAE: 0.6253634598854412
R2 : 0.19675135094369778
```

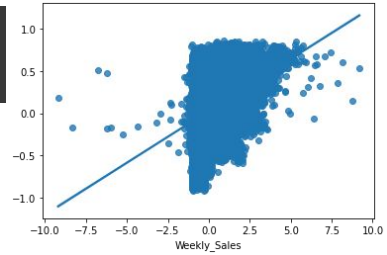


Linear Regression

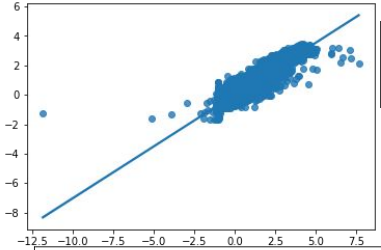



```
MSE: 0.8161764441536359
MAE: 0.6495531298298004
R2 : 0.22413646801991738
```

```
MSE: 0.79236680686
MAE: 0.6698936019286715
R2 : 0.1391413916810934
```

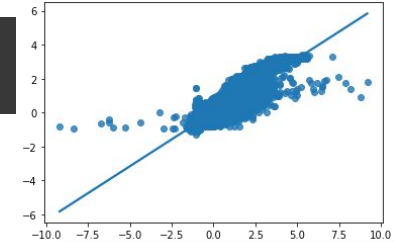


Ridge Regression

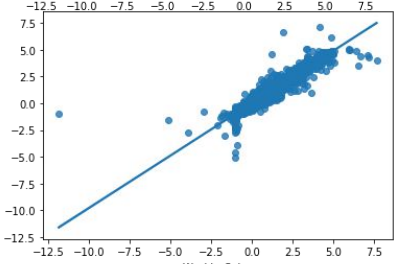



```
MSE: 0.20008958647788302
MAE: 0.31092388119724873
R2 : 0.8097933181125454
```

```
MSE: 0.2276660170837622
MAE: 0.3426508749719278
R2 : 0.7526546431129537
```

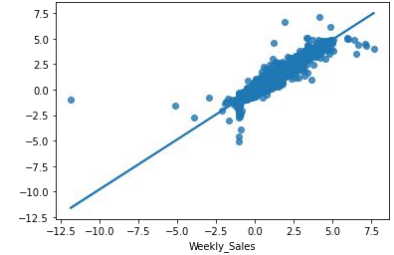


XGBoost




```
MSE: 0.03528361481900656
MAE: 0.08192329053644175
R2 : 0.9664591275445508
```

```
MSE: 0.03543457281792551
MAE: 0.08072807896945104
R2 : 0.9615024799394405
```



Random Forest Classifier

Performance Metrics

Models	Performance Matrix:Store 1_10	Performance Matrix: Store11_35
Linear Regression	MSE: 0.7880411373245544 MAE: 0.6246598472526079 R2 : 0.2508821045622597	MSE: 0.7393404224768703 MAE: 0.6253634598854412 R2 : 0.1967513509436977
Ridge Regression	MSE: 0.8161764441536359 MAE: 0.6495531298298004 R2 : 0.2241364680199173	MSE: 0.79236680686 MAE: 0.6698936019286715 R2 : 0.1391413916810934
XGBoost	MSE: 0.2000895864778830 MAE: 0.3109238811972487 R2 : 0.8097933181125454	MSE: 0.2276660170837622 MAE: 0.3426508749719278 R2 : 0.7526546431129537
Random Forest Regressor	MSE: 0.03528361481900656 MAE: 0.0819232905364417 R2 : 0.9664591275445506	MSE: 0.03129321660979816 MAE: 0.0804491397465661 R2 : 0.9660018129642668
ARIMA Model	MSE: 30.633598385018153 MAE: 4.902795855140299 R2 : -20.788963944926333	MSE: 1.8384723934543519 MAE: 0.9341963836686561 R2 : -0.8047079119326921

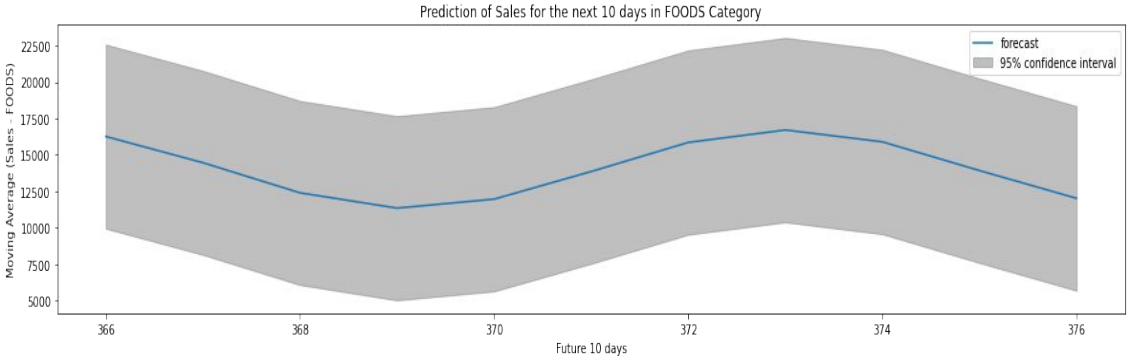
| Dataset - 2 ~ EDA & Feature Engineering

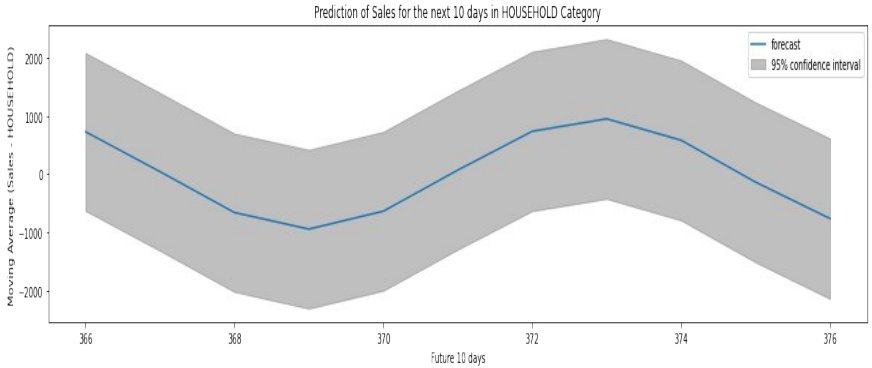
1. Data Visualization - Tableau
2. EDA - Sales Distribution across stores, areas w.r.t days, months & weeks
3. New Features - Weather and Median Income

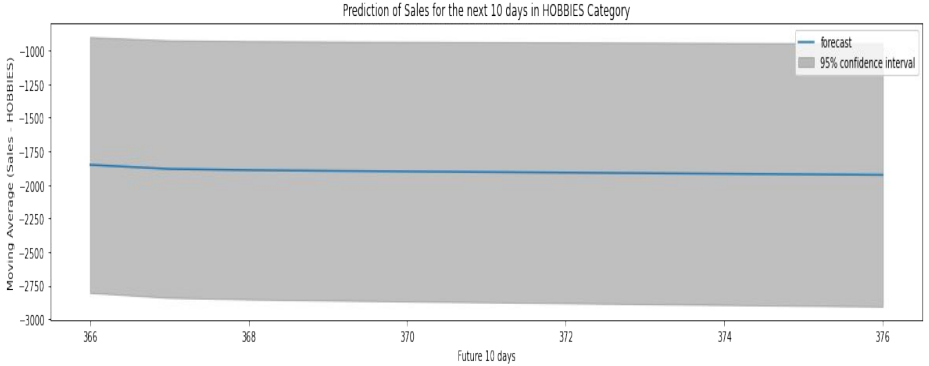
	d	cat_id	date	weekday	month	event_name_1	event_type_1	snap_CA	snap_TX	snap_WI	NAME	WT01	WT03	WT04	WT05	WT06	WT11	Med_Income	sales
0	d_1	FOODS	2011-01-29	Saturday	1	no_event	no_event	0	0	0	FREMONT, CA US	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	84840	1223
1	d_1	FOODS	2011-01-29	Saturday	1	no_event	no_event	0	0	0	FRISCO 1.9 N, TX US	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	77235	2788
2	d_1	FOODS	2011-01-29	Saturday	1	no_event	no_event	0	0	0	FRISCO, TX US	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	77235	2788
3	d_1	FOODS	2011-01-29	Saturday	1	no_event	no_event	0	0	0	LOS GATOS CALIFORNIA, CA US	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	118713	2193
4	d_1	FOODS	2011-01-29	Saturday	1	no_event	no_event	0	0	0	LOS GATOS, CA US	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	118713	2193

| ARIMA Model Implementation

- Time Series analysis
- Stationarity ~ ADF Test
- Predictions of 10 days further







Plots to show the sales predictions through ARIMA model for each category (Foods, Hobbies, Household)

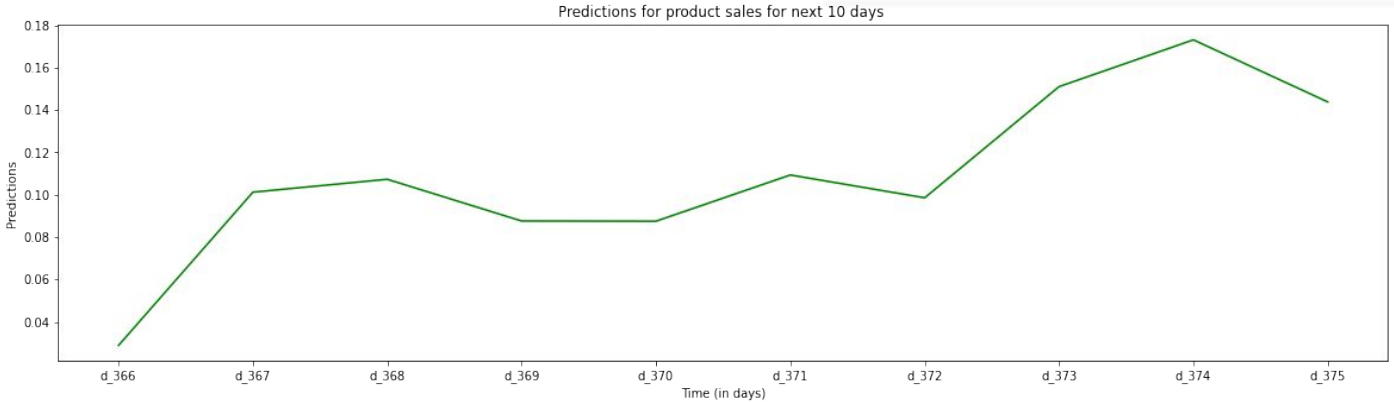
LSTM Model Implementation

- LSTM layer - 1, dense layer - 1
- RMSE score for LSTM model: 0.31

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 64)                21504

 dense (Dense)               (None, 19)                1235

=================================================================
Total params: 22,739
Trainable params: 22,739
Non-trainable params: 0
_________________________________________________________________
```

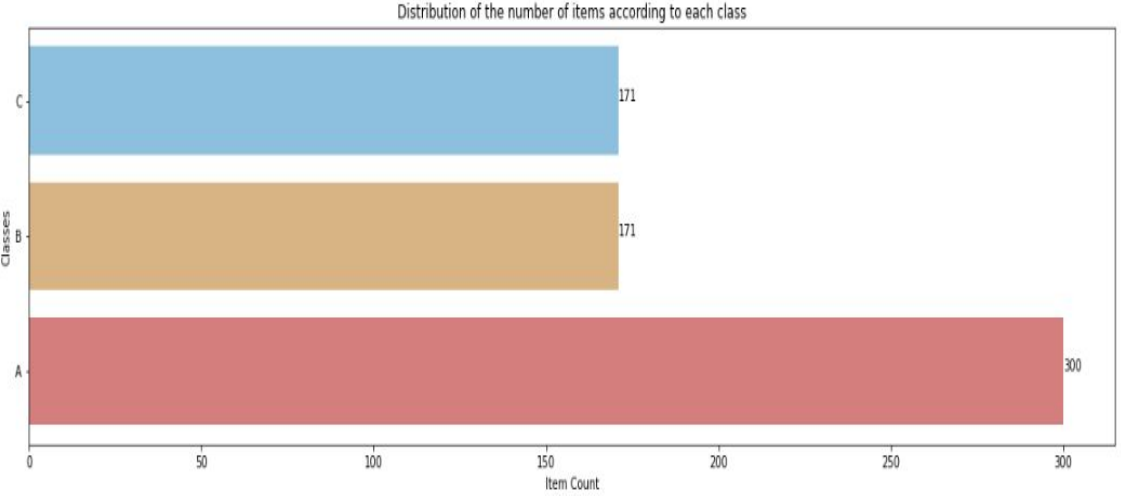


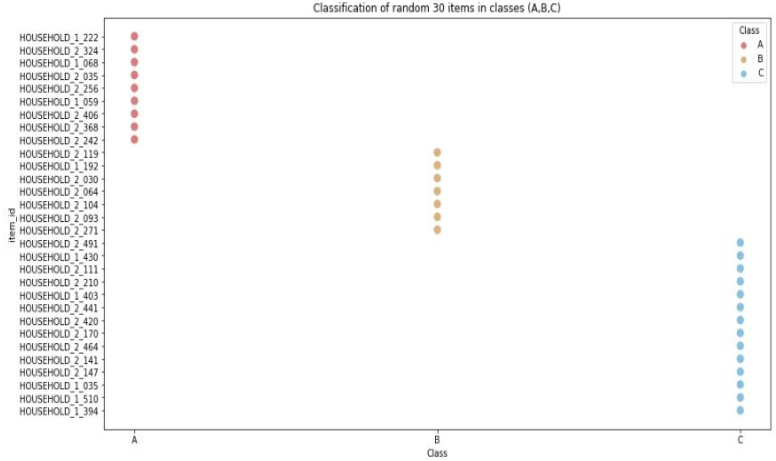
Plot to show the sales predictions through LSTM model for total sales

| ABC Analysis ~ Product Segmentation

- Classification of items in HOUSEHOLD Category
- Class A - 80% of the revenue, Class B - remaining 15% and Class C - the rest 5% revenue
- Items that constitute 80% of the total revenue are more than 20% of the total items (46.73% of the total items)
- Pareto Law does not apply

item_id	sell_price	sales	revenue	percentage	Class
HOUSEHOLD_1_272	9.97	11045	110118.65	1.76	A
HOUSEHOLD_1_535	6.97	12046	83960.62	1.34	A
HOUSEHOLD_1_053	14.97	5401	80852.97	1.30	A
HOUSEHOLD_1_537	15.97	3866	61740.02	0.99	A
HOUSEHOLD_1_177	7.97	7636	60858.92	0.97	A

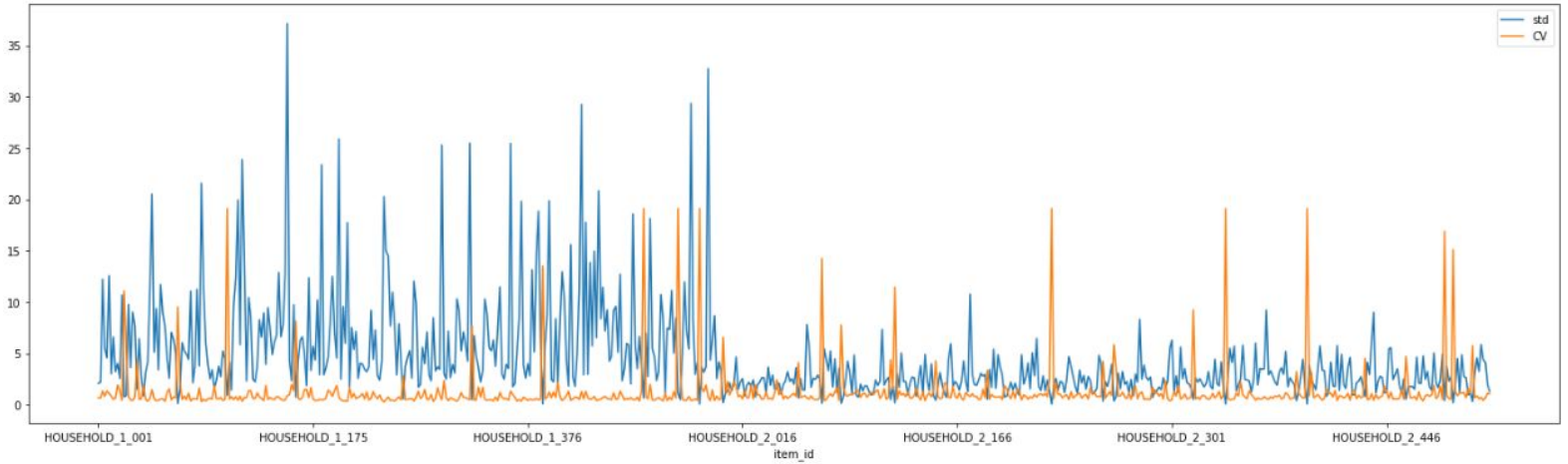




| Coefficient of Variation

- Variability in demands of items
- Some items tend to vary more than other items
- Sales in the range of about 125 to 150 deviate the most from the mean sales

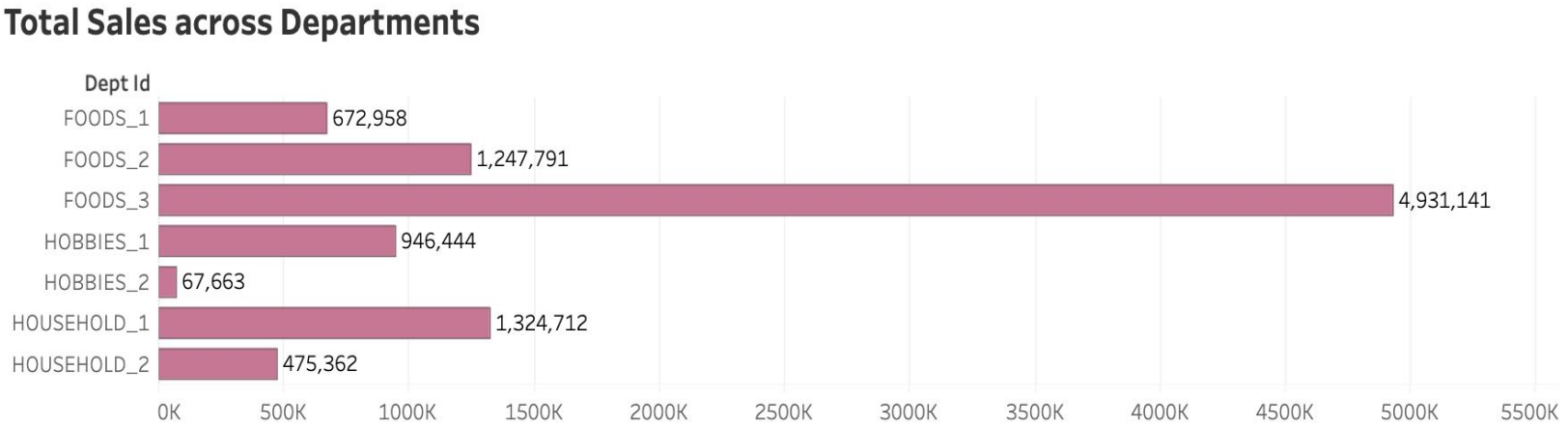
item_id	std	CV
HOUSEHOLD_1_001	2.063595	0.648761
HOUSEHOLD_1_002	2.171559	0.649156
HOUSEHOLD_1_004	12.215944	1.333379
HOUSEHOLD_1_005	5.464366	0.798756
HOUSEHOLD_1_006	4.543668	1.312056



Distribution of Standard Deviation and CoV over the HOUSEHOLD items

| Initiative ~ 1

- FOODS_3 has the highest sales
- Recommendation ~ increase in the restocking of these items
    - Regular stocking
    - Customers' demands will be met

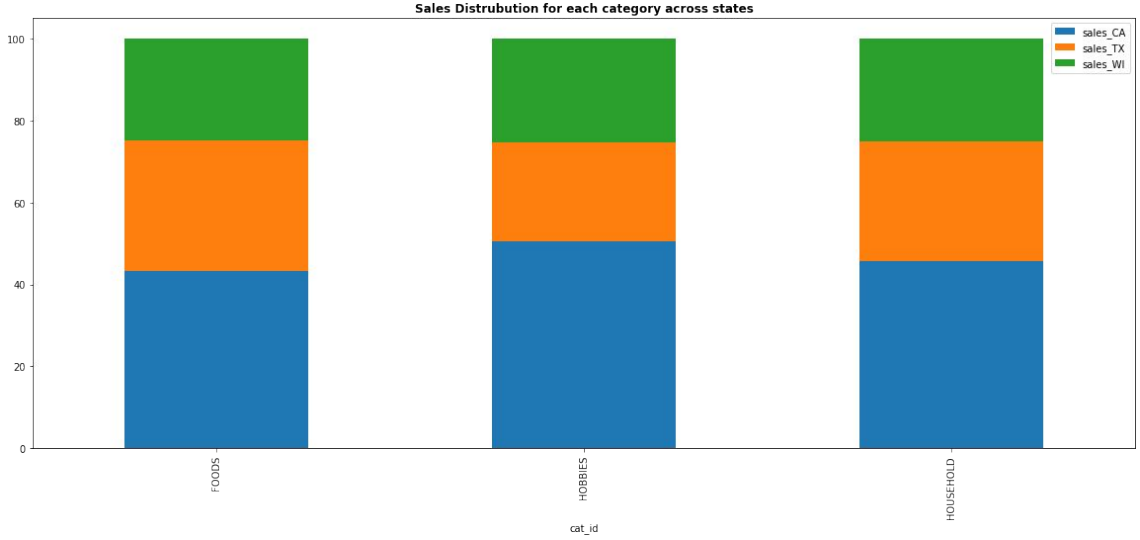


| Initiative ~ 2

- California constitutes 50% of the total sales in the HOBBIES category
- Recommendation ~ include more diverse items in this category (in the average and above average price range)
    - Enhance sales in HOBBIES
    - More profit gain in this category

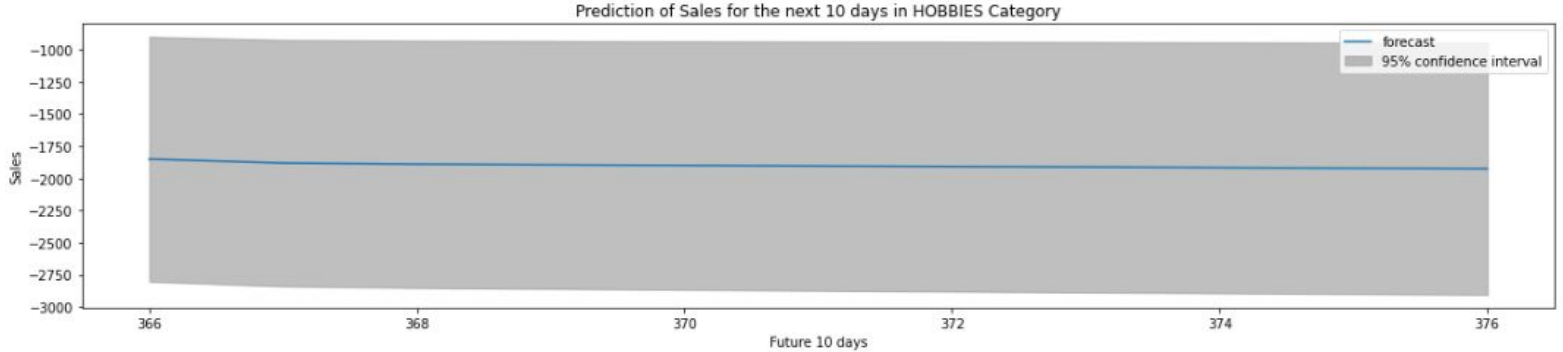
`data_in`

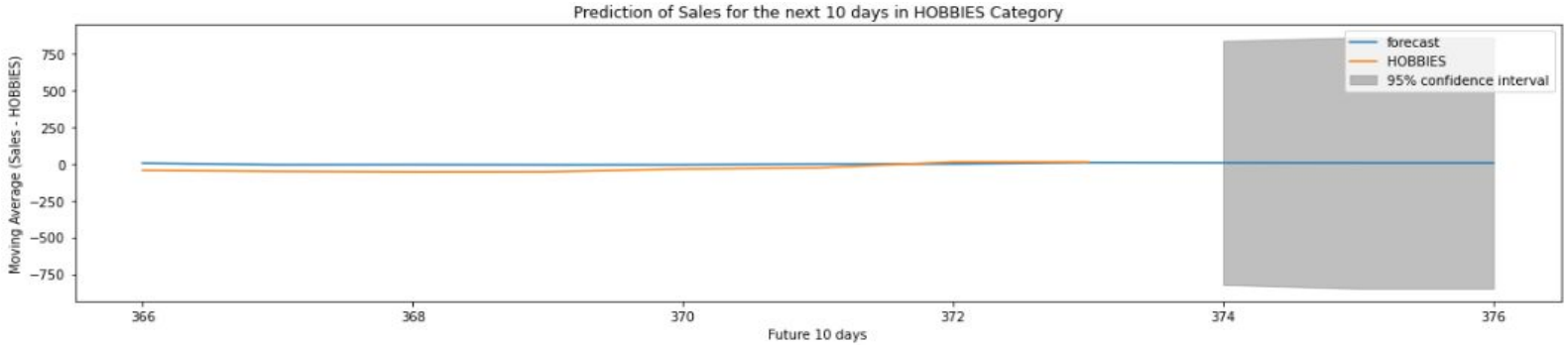
cat_id	sales_CA	sales_TX	sales_WI
FOODS	43.37492	31.85724	24.76784
HOBBIES	50.53875	24.09529	25.36596
HOUSEHOLD	45.71923	29.12053	25.16024



| Initiative ~ 2 Comparison

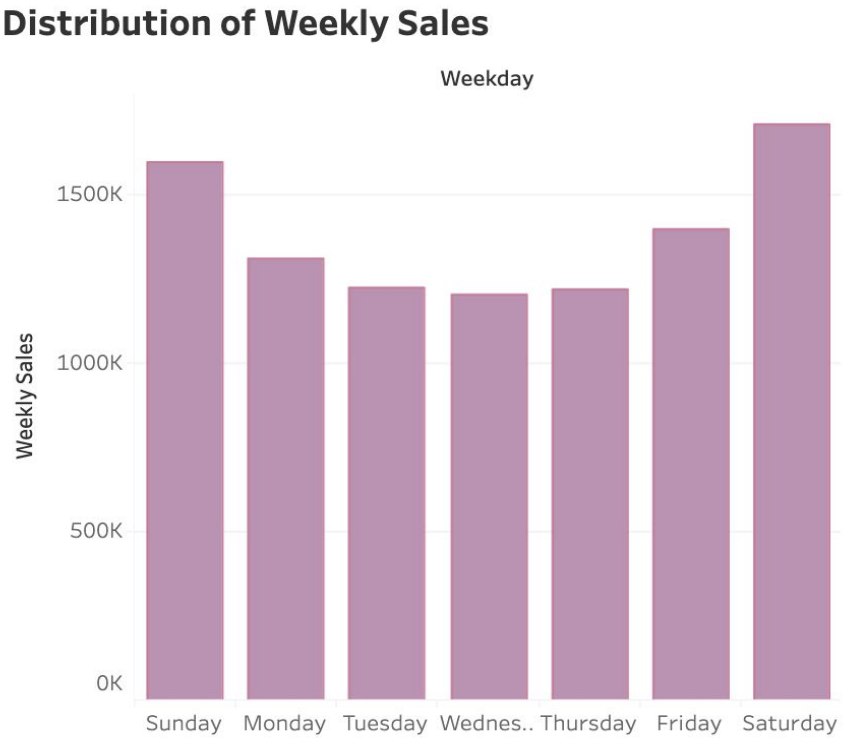
- Mutated a part of the original dataset and added a few more dummy items for the HOBBIES category
- Implemented the ARIMA model on the new series





| Initiative ~ 3

- Overall sales decrease on weekdays as compared to the weekends
- Recommendation ~
  - provision of free delivery service on an order of above $25 during weekdays
  - discount of 10% on the order consisting of top 20% items that are in demand



| Contribution

	Sprint1	Sprint2	Sprint3
Almee Christian	Tableau Dashboard	Feature Engineering,LSTM, performance matrix	ABC analysis,Initiatives
Harsha Bhargav	Q2 C,E	ADA Boost, CNN,Ensemble models	Coefficient of variance
Pragya Sharma	Tableau Dashboard	Data Preprocessing, EDA, Arima for all 3 categories	Product segmentation,Initiatives
Suraj Shah	Q2 A(EDA)	Arima, Ensemble models	Plots for ABC analysis, Initiatives
Taher Vora	Q2 B,D	Linear regression, XGBoost, Ridge regression, Random Forest Classifier.	Model deployment in Streamlit

Thank you!

Any Questions?