Usable AI : Project Selection Report

Team : Mitali Tavildar, Shefali Luley

- **Link to the original dataset:** https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset

- **Link to the Python notebook**: https://colab.research.google.com/drive/1EVwpCloUHRMlQgHg63ytQqZaj2m0Wq8G#scrollTo=70l2vzTo3kFY

Problem Statement:

The main objective of the project is to understand the different time series models and apply them in an appropriate context. The dataset that we have chosen contains details of purchase of a variety of grocery items over time & we aim to utilize this dataset to predict the demand for these different products in future. This will help manage inventory and increase efficiency for the distributors.

Data Description:

- The data which will be utilized throughout this project is of Groceries dataset.
- There are 38,765 rows in the dataset that contain purchase orders made by customers at grocery stores.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38765 entries, 0 to 38764
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Member_number    38765 non-null  int64 
 1   Date             38765 non-null  object
 2   itemDescription  38765 non-null  object
dtypes: int64(1), object(2)
memory usage: 908.7+ KB
```

Figure 1

Columns:

- Member_number: Unique id of the customer which is of the datatype ‘int’
- Date: Provides the date of purchase for the corresponding commodity. This is of the data type ‘object’

- itemDescription: It is the description/name of the commodity being purchased. This is in text format with the datatype ‘object’.

Observations from the data:

- The data we have covers commodities purchased for the year 2014 and 2015.
- There are a total of 38765 rows, and 3 columns. Out of which,there are a total 167 unique products in the itemDescription column.

Data preprocessing/Feature Engineering:

- The data was in the form of a csv, which using Pandas, we were able to import into a dataframe.
- We observed in the data description(Figure 1), that there are no null values.
- Since there are 167 unique values, and felt the need to categorize them into different food product types that they belong to.
    - We engineered a new column called ‘Category’ that broadly described the kind of the product.
    - For example, ‘chicken’, ‘beef’ and ‘sausage’ would be classified as ‘Meat’.

Summary statistics:

- The data present is for the items purchased between the dates ‘01-01-2014’and ‘31-10-2015
- We listed out the frequency of the categories of the products based on the count in Figure 2.
- The visual representation of the same can be seen in Figure 3. Through the heatmap,we observed that Dairy, Beverages and Vegetables are some of the categories with the highest bought products in the last 2 years.

Number	Category	Count
1	Dairy	5725
2	Beverages	5083
3	Vegetables	4572
4	Meat	4313
5	Sweets	4064
6	Bakery	3417
7	Fruits	3167
8	Condiments	2910

9	Kitchen supplies	2006
10	Misc products	1913

Figure 2

Figure 3

- Figure 4 represents the count of each product category in the year 2014 and 2015 separately. This plot provides an overview of the demand of each category year wise.

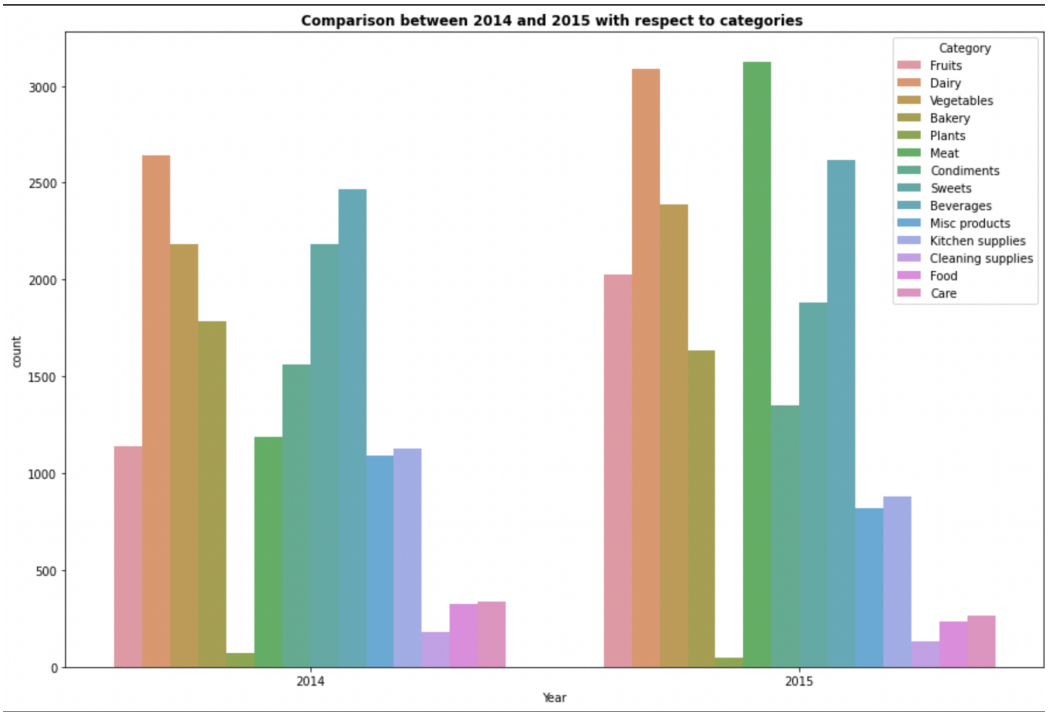

Figure 4

Basic visualizations for data analysis:

- In order to forecast the demand for each category, we needed to understand and analyze the demand of the historical data.
- We began analyzing if there exists any trend in the demand pattern from a very high i.e, yearly level to a very low, i.e daily basis.
- Figure 4 clearly shows the yearly demand for each category. We observe the count for meat,dairy having highest demands.
- In Figure 5, we then drilled down to analyzing monthly trends for demand for each country. We observe a noticeable difference in the range of count for each category.

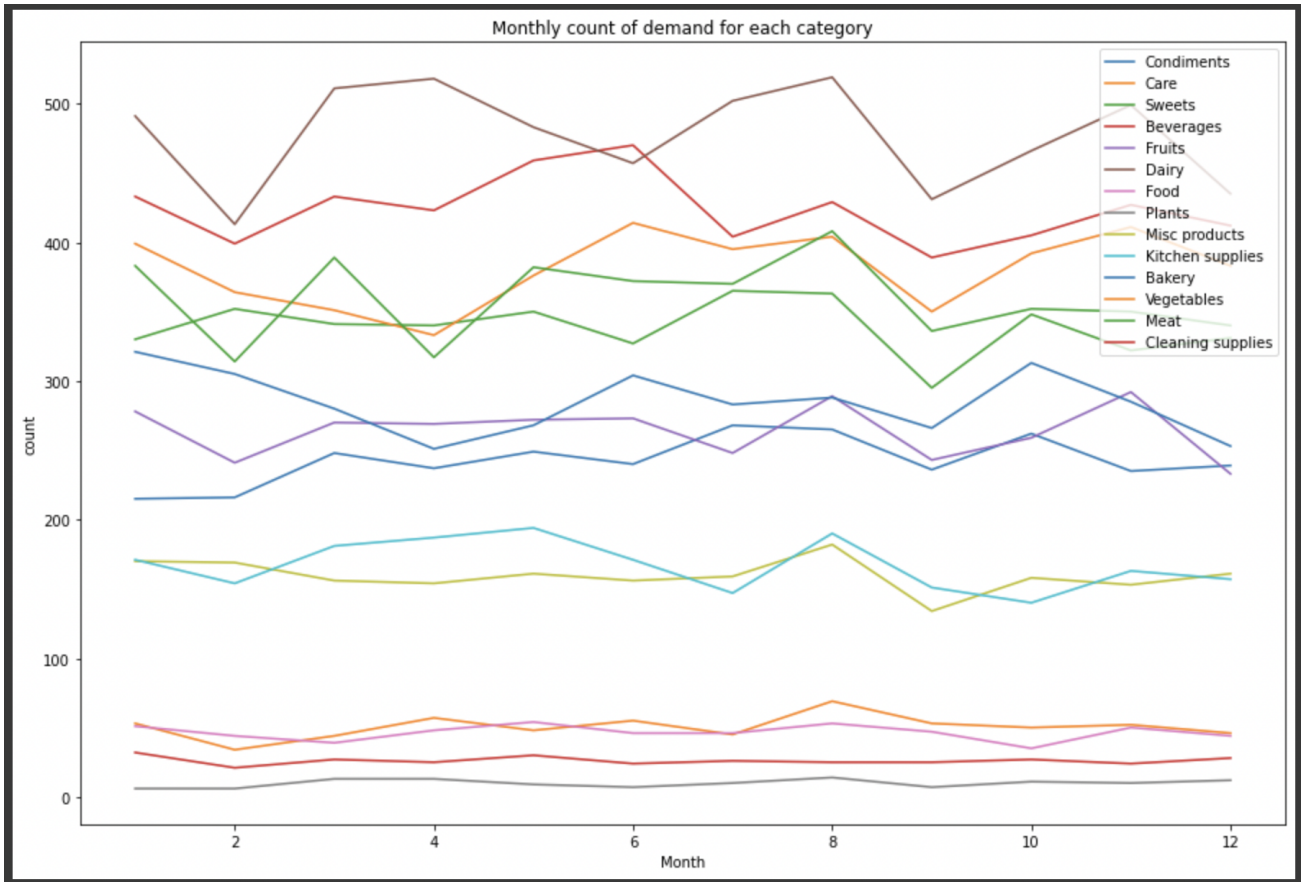

Figure 5

- In Figure 6, we drilled further down and analyzed a plot for demand over 2 years for each category.
- Figure 6a represents these demands for the year 2014, similarly Figure 6b represents these demands for the year 2015.
- We observe an overlap in certain categories, and notice the demand trends to be similar for these.
- In figure 7,The graph above indicates a consistent pattern in daily sales from 2014 to 2015, with the exception of a dip in January 2015.
- The cause of this dip is uncertain and may be attributed to various factors such as inflation, among other factors.

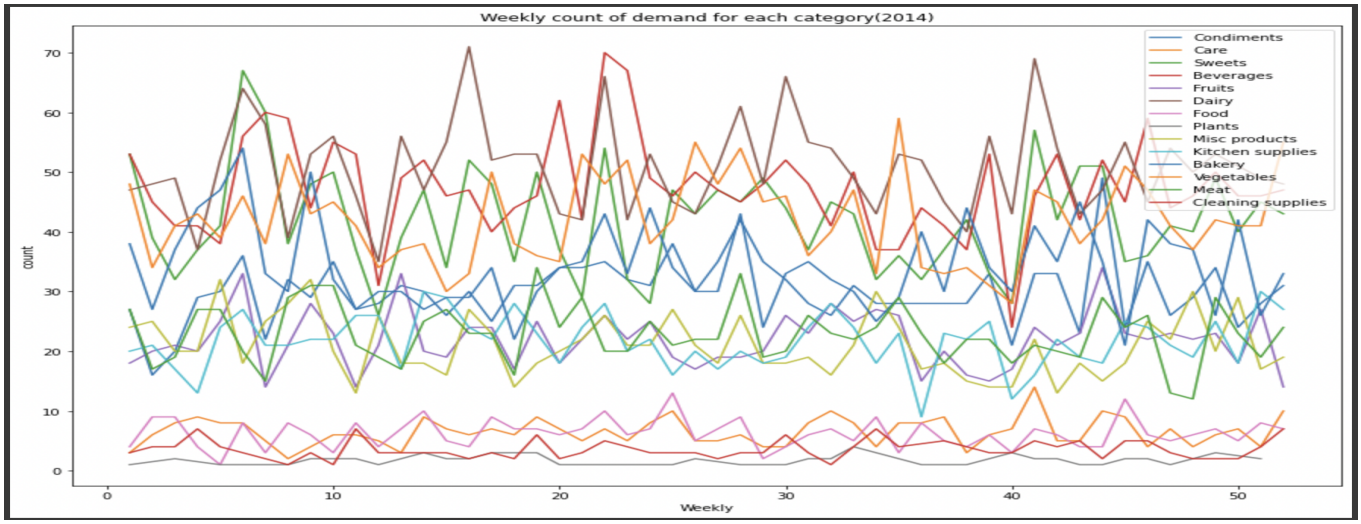

Figure 6a

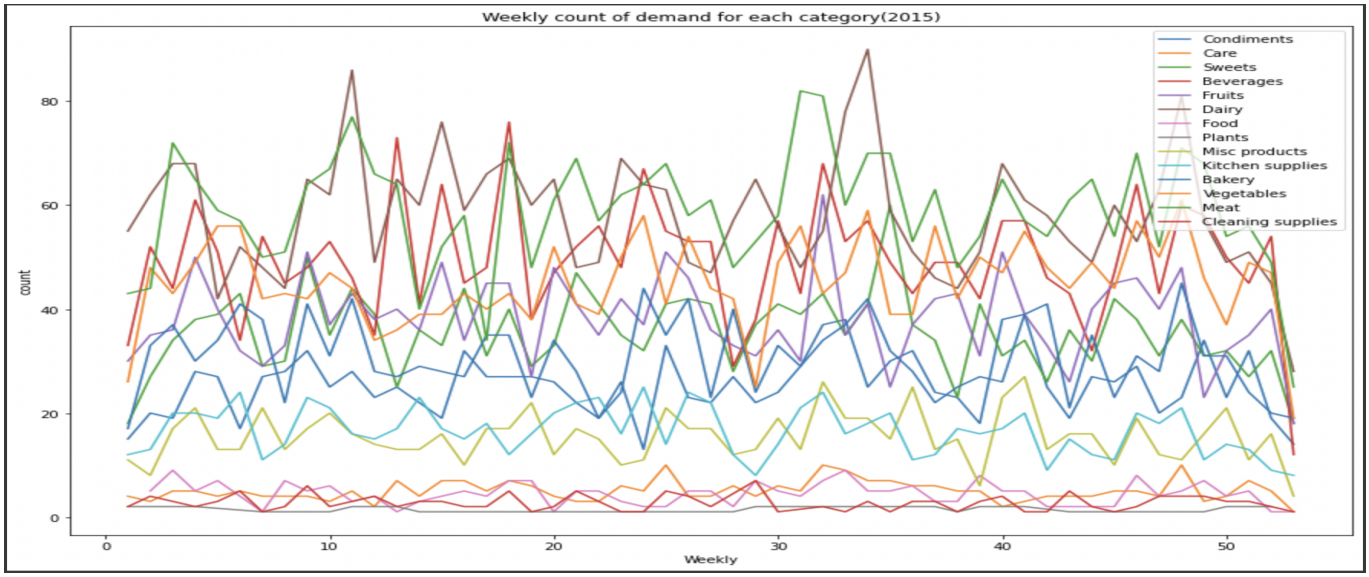

Figure 6b

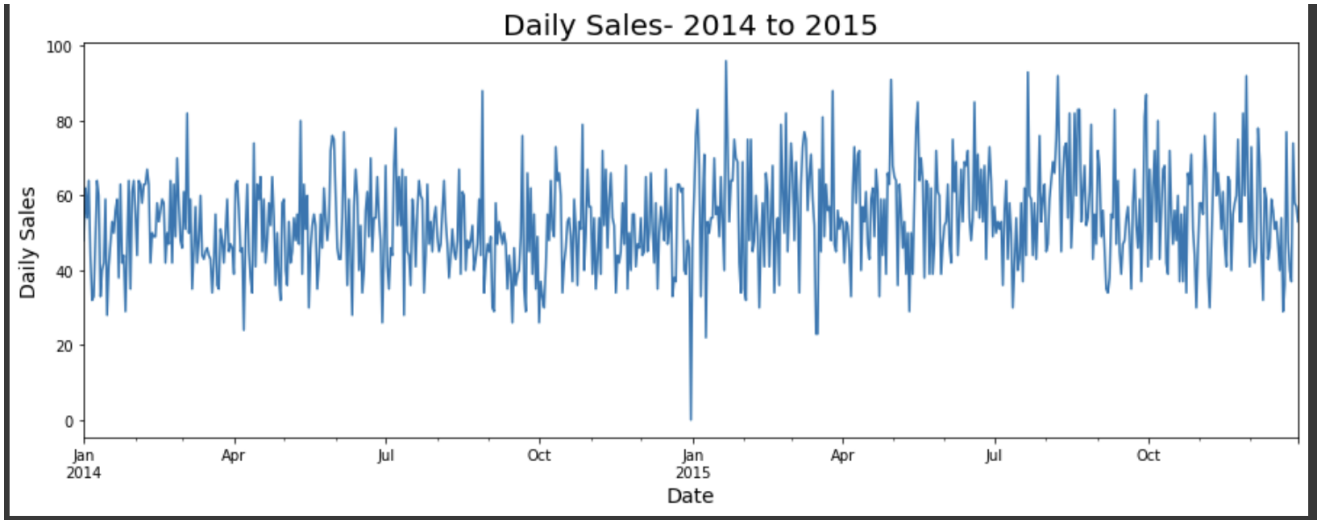

Figure 7

Result:

- In all the above graphs and statistical analysis, where we are trying to look into variation in demand of different categories over months, weeks and on a day to daily basis.
- It can observed that there are few factors and categories which have a significant amount of dominance on the whole patterns and trends

Research and Design Questions:

- Are there seasonal or trend-based patterns in customers' purchasing behavior that can be identified and accounted for in a forecasting model?
- What is the most appropriate forecasting method for this type of data (e.g. time series analysis, regression analysis, machine learning algorithms), and how can its accuracy be evaluated?
- How can data preprocessing techniques such as feature scaling, outlier detection, or missing value imputation improve the accuracy of a forecasting model?
- What are the trade-offs between a simpler, more interpretable forecasting model and a more complex, potentially more accurate model, and how can these trade-offs be optimized for the specific business problem at hand?
- How frequently should the forecasting model be updated to account for changes in customers' purchasing behavior or external factors that may impact demand?
- Can clustering or segmentation techniques be used to group customers based on similar purchasing behavior and improve the accuracy of the forecasting model for each group?
- How can uncertainty in the forecasting results be quantified and communicated to stakeholders, and what is the best approach for making decisions based on uncertain forecasts?

- How can the forecasting model be integrated with other business processes such as inventory management, supply chain planning, or marketing campaigns to optimize business performance?