MIDTERM PROJECT REPORT

Submitted by: Mitali Tavildar
Shefali Luley

DATA COLLECTION:

- **Where did the data come from?**
  The data comes from Kaggle dataset:
  https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset
  The data contains details of purchase of a variety of grocery items over time.

- **How was the data collected?**
  The data was gathered from retail establishments that possess records of customers' purchase orders in their grocery stores.

- **Did it come from a poll/survey, was it scrapped (from where? Is there code?),was it hand collected?**
  The data was not obtained through any polling or surveying methods; rather, it was acquired from grocery stores.

- **Can I verify where it came from (e.g. from a website)?**
  There is no specific website or database mentioned except data was collected from grocery stores which contains purchase orders.

- **Could I recollect it, or is that impossible (e.g. proprietary data, or something has changed since it was collected)?**
  The retrieval of data is feasible, as grocery stores have the capability to store transactional records of customer purchases.

- **Double checking: was any aspect of the data collected manually?**
  Perhaps, the data entry of purchases could have been done manually, but there is no mention of it in the data description.

DATA MANAGEMENT

- **How did you clean and pre- process the data?**
  We tried to identify any null values in the data, but there weren't any. The data was stored in CSV format, which we were able to import into a Pandas dataframe for further analysis.

- **What recoding of the variables was done after the data was collected?**
  Since there are 167 unique values, we 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’.

- **Is any other data merged in?**
  There is no other data merged, we only added columns to divide into categories.

- **Was any of the data manipulated manually after it was collected?**
  As the dataset was devoid of any missing values, the data cleaning phase was omitted during the analysis. However, the analysis did encompass the creation of new columns derived from the original data in order to gain a more comprehensive understanding and generate meaningful insights.

- **If yes, is there a record of these changes?**
  All the modifications made to the dataset can be viewed in the Python notebook, which includes detailed comments and headings to provide clarity and context to the changes.

ANALYSIS: Provide a justification for the methods you applied to your analysis

- **What is the analysis done here?**
  We began by examining the temporal scope of the dataset. Afterwards, we classified the items contained within it and evaluated the frequency of their purchase. Using this information, we generated a summary of the demand for each category, taking into consideration the year, month, and week in which the purchases occurred.

- **What were the decisions made?Choice of method? Did you try anything else?**
  According to our problem statement our clear choice was to go ahead with forecasting. We researched different factors to be considered for forecasting prediction of each category of item.In order to analyze time series data, we experimented on the seasonality factor of it. Using ‘seasonal_decompose’ model from the statsmodels library, we selected a multiplicative model out of the two options:
    a. Additive model
    b. Multiplicative model

  since it would appear the amplitude of the cycles is increasing with time. Which made sense with respect to our dataset.
  From the resulting graphs, we interpreted the following:
    a. trend — the general direction of the series over 24 months(2 years)
    a. seasonality — a distinct, repeating pattern observed in regular intervals due to various seasonal factors, which was monthly in our case
    b. residual — the irregular component consisting of the fluctuations in the time series after removing the previous components

- **Is there any data subsetting?**
  There's data categorizing but not data subsetting. We will be utilizing the top 3 categories which are mostly in demand that are:
  1. Beverages
  2. Dairy
  3. Vegetables

- **Is there any subgroup analysis? How large are the subgroups?**
  We created a total of 14 categories along with a calculation for demand of each category. For further analysis, we calculated the statistics of monthly, year and weekly demand of each category. We have also created a countplot stating the change in patterns between the two years of data.
  We started with a general trend analysis of all the data items, followed by trend analysis of the top 3 categories as mentioned. This analysis was done using the 'seasonal_decompose' from statsmodel library.

- **How many tests are run? What kinds of tests?**
  We are implementing the concepts of time series analysis on an experimental basis by trying to fit our data to certain algorithms such as ARIMA,ARMA,SARIMA that will give us the best predictions. We plan to perform cross validation or scenario analysis to overcome and analyze the errors.

- **Is there uncertainty in the estimates?**
  Yes, We expect some of the models not to fully capture the complexity or some extremely sophisticated or complex for our data.
  Based on the initial analysis, we found the trend graphs for the generalized items and specific top 3 categories to be extremely varied. Hence, we plan to implement the best fit algorithms and test the estimates accordingly.

- **How do you show that uncertainty (e.g. confidence intervals, standard errors, hypothesis tests)?**
  We plan to plot our train test and validation error graphs to identify this performance uncertainty and act on it.According to our experiments , we observed the trend prediction for the future is highly dependent on the initial trend of our original data.
  We tested our SARIMA model implementation of the original dataset and are currently working on error metric evaluation.

- **Is there anything you're unsure about?**
  We have implemented the SARIMA model for future sales but we are unsure if this is the perfect algorithm to be used for our use case. We plan to experiment with a few more forecasting models to predict the demand for different categories of product in our data.

ARGUMENT: Provide the key core argument of your data interpretation

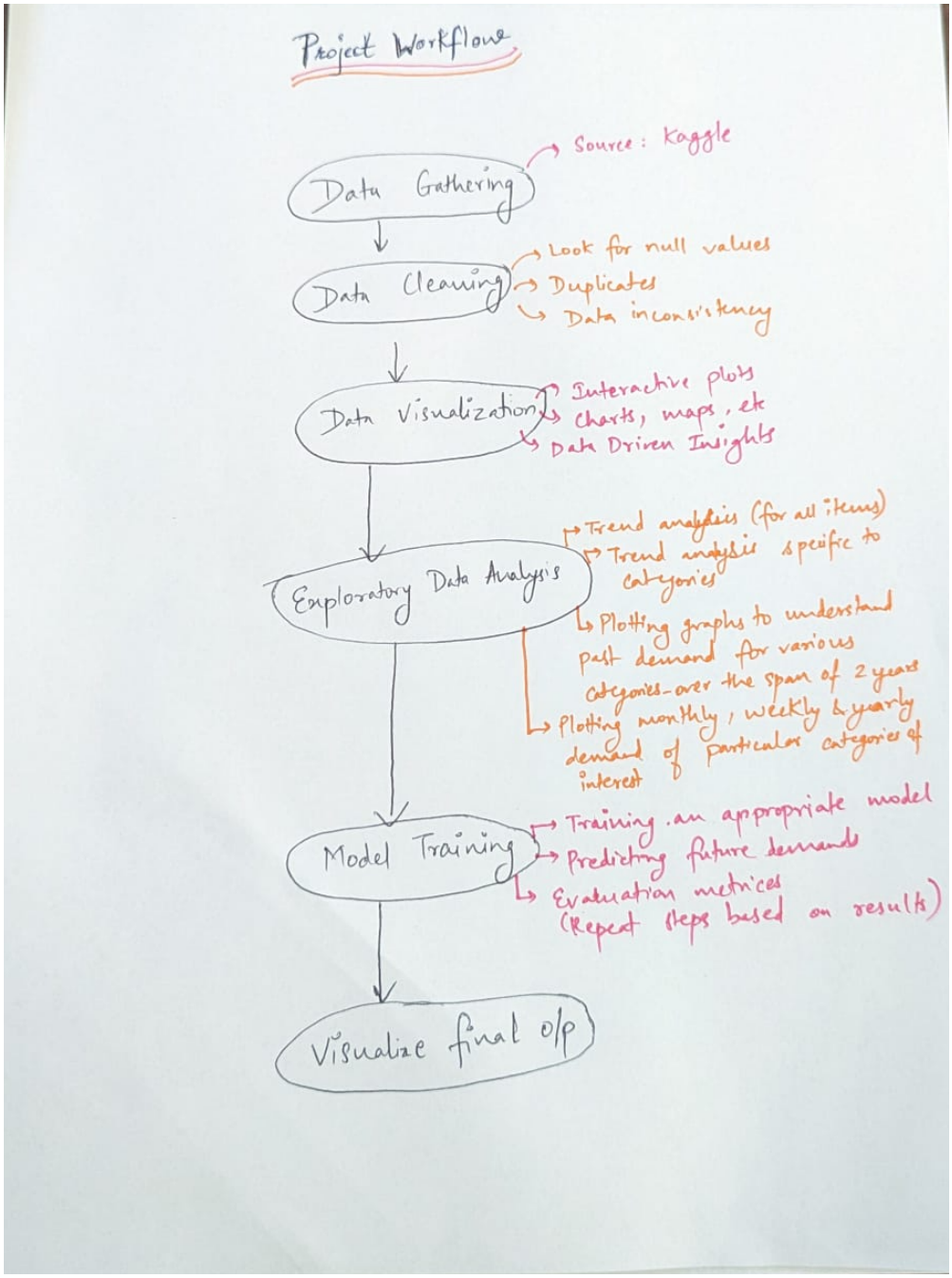
- **What's the core argument being made here?**
  We are trying to identify the general pattern for demand of the top 3 categories, using which we plan to predict the future demand of the same.
  Our main assertion is that the demand for these categories has consistently been the highest over the course of two years. Therefore, we put forward the argument that the demand for these categories will also remain the highest in the future.

- **What is the evidence provided to answer it?**
  Based on the trend analysis, and resultant plots, it is apparent from particularly the trend graph that the demand would remain highest. Given that these are essential grocery items, it is highly likely that customers will purchase them in future as well.

- **Is it a causal argument?**
  No, it is not a causal argument.

- **If yes, what is the evidence that speaks to the causality?**
  NA

[Design]



STAKEHOLDER ANALYSIS: Define your target stakeholder groups and their user needs

- **Who cares about this data? Who is the specific user group you are targeting?**
  The target audience for this information would likely be individuals or organizations interested in the retail industry, particularly those involved in market research, investment analysis, and strategic planning. This could include analysts, investors, consultants, and business owners who want to stay informed about the major players and trends in the retail sector, as well as understand the competitive landscape and potential opportunities for growth and investment.

- **Who does your data affect? What are the relevant user characteristics: ability/disability,gender, background, age, attitude to computers? What level of computer experience will they have (and is this important): novice, expert, casual, frequent?**
  The data majorly affect the consumer and the retailer. This prediction would be relevant to retailers to optimize their inventory and be prepared for seasonal lows and high prices of products. Retailers can also utilize the user data to understand the buying pattern of each user to target advertisements hand picked as per their choices. This can also benefit the retailers to manage their finances accordingly and optimize their budget.

- **What needs do your target population have?**
  Our target population is “The Retailers” . and their basic need is to generate good revenue and satisfy customers' needs. So when we talk about customers' needs, the retailers require to analyze the data , not in terms of the regular customers but also in terms of the customer churns.It means the retailers might need to analyze what are the lay back or items missing in the inventory due to which there’s less customer frequency.

- **What are the underlying drivers and motivations of this population?**
  One of the most important underlying drivers of the retail industry is the consumer demand. “Groceries” are the necessity to sustain life hence its important to keep up with change and their preferences. The second one would be competition, competition in a sense where there’s competition between two retailers, where most of them are inclined to attract consumers and retain them. Methods like offering discounts, incentives are few of them.One more observation which we had while doing our studies was the term “Sustainability” , initially it wasn't a big factor in terms of groceries but now it is. Consumers are inclined more towards a healthy lifestyle and are concerned about environmental impact on their food choices, due to which industries have started reducing waste and more.

- **How can your stakeholders use these insights?**
  Retailers and stakeholders can use these insights since predictive models can be used to forecast demand for different products, which can help in making the optimal use of resources.This will also help in making decisions such as whether overstocking is required or not as per customer ratio.It can also be utilized to manage customer segmentation, where predictive models can be used to improve the effectiveness of marketing campaigns based on their behavior and preferences.

  CONTEXT: Define the use case and workflow of data interaction

  • How will the stakeholders engage with the data?
  For stakeholders who are interested in conducting market research such as surveys,etc can use this data to analyze trends and patterns in the retail industry. This will also give them a benefit where they can develop further strategies to stay ahead of their

competitiors.As we said, “ develop strategies”, stakeholders can also consider investing in other retail industries to evaluate financial performance and growth potential of other companies like Target, Kroger etc. They can also identify the potential areas of growth or expansion ,evaluate the feasibility of entering new markets

• How will information be shared and viewed?

The sharing of information solely depends on the preference of the retailers/shareholders. Some of the information might be shared in forms of reports, presentations, some data visualizations showcasing the trends and patterns , in terms of differences in every year, month and more. Another way could be accessing a remote/ online available database that provides access to chunks of information from that company where you can provide stakeholders to manipulate and analyze the data.

• Are there any privacy or trust issues that need to be considered?

Data Privacy and trust issues are a crucial part in any industry . Various Data protection measures such as access controls, encryptions should be implemented. Keeping personal information anonymous is also important since their chances of cyber crimes with someone’s credentials.It is critical to safeguards one’s rights and privacy as well establishing trust amongst stakeholders.

• What requirements come about because of the context of use?

The requirements that come from the context of use will specifically depend on the needs and objectives of stakeholders. To begin with , Data Quality would be the most important amongst others since up-to-date, cleaned and validated data might be required to ensure that its accurate.Another can be accessibility, where the stakeholders may require easy access to the data , to manipulate and download it in various formats,etc.

VISUALIZATION: Present data visually

• What effects are you trying to achieve by showing this data to your target stakeholder groups?

The effects which we are trying to achieve by showcasing the results from the Exploratory Data Analysis and the time series analysis to our stakeholders are basically communicating insights and identifying opportunities and challenges along with forecasting future trends and seasonal changes. Data Visualization will help the stakeholders to understand the data and will help in making informed decisions.On the other hand, time series analysis here reveals patterns and trends over months which will help stakeholder groups to plan further actions.

• What are the effective data representations that could captivate your audience?

The effective data representation that could captivate the audience depends on the data and audience preferences. Most of them can be represented in the form of visualizations, interactive dashboards and graphs , data storytelling and more.Visualizations can be in the form of graphs, charts, maps showcasing networks (graph theory) where one can interact and engage with it more for better understanding.

• How can you use design to clearly convey your message?
To use design to clearly communicate one needs to understand what type of audience you are trying to reach , is it a technical audience or a some non- technical part of an industry. Accordingly one needs to use visuals to explain complex information in an effective manner. Another thing would be to take feedback from the audience to progress every time you are using design to convey your message which will fulfill consumer needs.

ETHICS: Consider possible unintended and negative consequences of using your data

• What values are relevant to your design?
The two most relevant values would be Data Privacy and Data Quality. Because we are dealing with user information including demographics and user choices, there needs to be a layer of security added such that user information is not misused.Data Quality is important because the whole trend analysis and prediction is based on the actual demand which needs to be right/true.

• How does your design portray your values?
The visualizations that we plan to implement would not include user demographics or such. It will only include the choices that the user has made,based on which different visualizations would be developed, thereby abiding Data Privacy.
We would try to stick to the original data received instead of modifying it to keep it true and protecting the original quality of it.

• What are some of the ethical and unethical ways of using your data?
Ethical ways as mentioned would be protecting the user demographics and other information.And on the other hand,unethical ways would be exposing user data such as user demographics,payment method,payment details,etc**.**

• Does your design disproportionally affect underserved, marginalized, low-resourced, and underrepresented populations?
This design would be generic to all the retailers, which is our target audience and is not expected to disproportionally affect underserved, marginalized, low-resourced, and underrepresented populations.

• Does your design impact the world's environment, resources, and climate?

Our design is predicting the future trend for grocery items that are frequently bought by users, and this design could only indirectly affect the world's resources.

• Are there ways to accomplish your personal and organization's mission and values while promoting positive change in the society?
Yes. Since we are working on trend analysis for grocery items, we can emphasize on analysis in terms of sustainability, inclination towards healthy food choices and reduction in waste material,thereby considering impact on environment. Based on the output, we could incentivize the audience in buying more sustainable items, thereby promoting positive change in the society**.**

References:

[1]https://towardsdatascience.com/finding-seasonal-trends-in-time-series-data-with-python-ce10c37aa861
[2]https://towardsdatascience.com/time-series-diy-seasonal-decomposition-f0b469afed44
[3]https://medium.com/towards-data-science/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b
[4]https://medium.com/towards-data-science/time-series-problems-simply-explained-as-fast-food-combo-meals-70c6eb9bdef