OliverWyman

HackAtOW

# HACK AT OW

Sustainable Promotions
First Challenge: Understanding Promotions

17th February 2023

A business of Marsh McLennan

# UNDERSTANDING PROMOTIONS IN A RETAILER IN ASIAN HEALTH AND BEAUTY

## Context

- Oliver Wyman have been engaged by an Asian health and beauty retailer to help design a set of promotions in line with their existing promotional offering and long-term sustainability goals
- The retailer has an online and offline (in store) presence in multiple countries across Asia but we have been engaged to focus on the Malaysian business unit and their offline promotions
- They have also asked to the team to focus on a subset of the products within a specific product category
- To support the Oliver Wyman team, the client has shared multiple datasets for us to understand the current performance of their products and the promotions they offer as they want to move towards a more data-driven approach. Previously individual category managers would rely on industry expertise to set up promotions and campaigns but were unable to review the effectiveness of the promotions coherently across the business

## Your task

- You are working with the new project team and as the data and analytics consultant on the project, you have been asked by the senior members of the team to assess the data shared by the client, conduct exploratory data analysis and build a simple model to predict product baseline sales
- Details of the tasks are in the following slides where it is expected that the material can be read and understood in a stand-alone context
- For any slides created, these should be clearly organised into an overall structure and submitted as a single PowerPoint file
- For any code used, this should be well commented, legible and submitted with clear documentation on how the code was used

# TASK 1: CLEANING THE DATA

## Context

- The local business teams shared the transaction data for their entire transaction data history. Unfortunately, some of the business teams use manual steps to create the transaction data, which impacts the overall quality of the data
- Additional details on the datasets provided are available in *Appendix A*
- Look at the data provided and prepare it so that it can be used for further analysis/modeling
  - What data quality issues exist in each dataset? Are values in line with expectations? Were there challenges in joining the datasets?
- The leadership team is also interested in the overall quality of the data and would like an assessment of what things can be improved to make the data more useful as the organization looks to become a more data-driven organization. What recommendations for improving their existing data can you think of?

## Your task

- Prepare up to four slides detailing the datasets received from the client, the data quality issues and subsequent remediation steps needed to build a clean dataset, and recommendations for improvements for their existing data
- Prepare one slide containing the following summary statistics for your final dataset over the full time period and the steps conducted to create the final dataset
  - Total sum of all sales, volume and discount
  - Sum of Discount by PromoMechanic
  - Sum of Sales, Volume and Discount by Online and Offline Stores
  - Sum of Sales, Volume and Discount for StoreKey = 2071
  - Sum of Sales, Volume and Discount for ProductCategory = Category AC
  - Sum of Sales, Volume and Discount for ProductKey = 49489

# TASK 2: UNDERSTANDING PRODUCT CATEGORY & PRODUCT PERFORMANCE

## Context

- Since the data has only recently become available, the client's leadership team would like to obtain data profiles for their product categories and leading products within the category
- For the next meeting with the client, concise deep dives should be provided, which will also serve as the basis for further deep dives that the client will conduct (please note that the client will also take your codes as a starting point and therefore, it is important to make the code very dynamic so that deep dives for other products can be easily performed)

## Your task

- Prepare two slides detailing the overall performance of the product category provided
  - This may involve consideration of dimensions such as online and offline sales, regions and product sub-categories, and trends over time
  - Are there any external factors which may influence long term trends? (e.g., seasons, holidays, natural disasters)
  - A dataset containing the Consumer Price Index for Malaysia has been provided to support this
- Prepare four (one per product) additional deep dive slides for products 49340, 49341, 49333 and 49329
  - Do you recognize similar trends when you compare them with the trends of the product category to which the products belong? Are there any significant changes in sales and discount, and why might this be?

# TASK 3: MODELLING PRODUCT PERFORMANCE

## Context

- The client wants to be able to understand the performance of products under promotion with a more quantitative measure
- In order to do this, we need to first understand the performance of products without promotion e.g., what is the baseline of sales without promotions, and then measure how effective promotions are on top of this baseline
- We will need to train a model to predict a baseline of sales without promotional activity. During this process, you may need to consider
  - If or how to use days where a promotion is active in the training data
  - What other factors may influence consumer behavior beyond promotion (e.g., weekends, seasons, holidays, natural disasters) and how they can be appropriately captured as factors in the model (binary, numerical, categorical etc.)
  - How to capture long-term trends connected to the product
- A common metric to measure promotion effectiveness for a given product is elasticity. This relates the additional uplift in sales gained to the discount given away and the client is interested in using it as a consistent promotion performance metric
- **Due to known data quality issues and variation in selling prices across stores and regions, the client and Oliver Wyman team have aligned on the definition that a product will be considered on promotion for a given day across offline stores if the sales discount is greater than 5% of the retail full price sales for the same day**

## Your task

- Create a promotional binary flag to determine whether a product is on promotion using the aligned definition for a given day across all offline stores. Where a discount is observed but below this thresholds, adjust the discount and actual sales values to align with the retail full price sales
- Using the additional information provided in Appendix B, train a machine learning model on the appropriate factors to predict the baseline sales for the following products **for all offline stores**: 49340, 49341, 49333 and 49329
- Using your model, predict the baseline sales for the four products over the full time period
- During this same time period, calculate the elasticity of the product across all promotions where we define
  - $Elasticity = -\frac{Sales\ Uplift}{Sales\ Discount}$, where Sales Uplift is the difference between the predicted Sales Baseline and Actual Sales
- Prepare up to two slides on the modelling and testing approach followed and how this aligns to best practice (e.g., factor selection, train/test samples, performance metrics). Since the client will use your proposed approach for further modeling tasks, emphasize what needs to be considered at each step
- Prepare four slides (one per product) detailing the predicted baseline sales compared to the actual observed sales and discount *(see slide 18 for an example visual)*, any model performance metrics, and the calculated elasticity
- Submit the code used to train and test the model including detailed model parameters and hyperparameters used

# SUBMISSION GUIDANCE

Overall, two main deliverables are expected from the candidates

1

**Notebooks or code-files from the candidates**

- For each section at least one notebook or code-file is expected
- Candidates must share codes in a zip file and code will be expected to run directly from raw datasets. This is a requirement for challenge submission
- Candidates are allowed to submit in Python (strongly encouraged) or R

2

**Slides**

- Slides must be shared in a single PowerPoint file in English language

Deadline to submit the deliverable is 12pm GMT on Sunday 5th March. Deliverables should be submitted by email to ow.europe.hackatow@oliverwyman.com

# A.

## DATASETS

# PROMOTION (HACKATHON_DIMPROMOTION_SAN_VSHARED.CSV)

This dataset details the promotions which were active and the range of dates they were considered active for

| Column | Description |
|---|---|
| **PromotionKey** | Unique Promotion identifier |
| **PromoMechanic** | Brief description of the promotion type |
| **PromotionStartDate** | Promotion start date |
| **PromotionEndDate** | Promotion end date |

# PRODUCT (HACKATHON_DIMPRODUCT_SAN_VSHARED.CSV)

This dataset details the unique products within the product category of interest, their sub-category and descriptive information including brand and supplier

| Column | Description |
|---|---|
| **ProductKey** | Product identifier |
| **BrandKey** | Brand identifier |
| **SupplierKey** | Supplier identifier |
| **ProductCategory_Lvl1** | Main category to which the product belongs |
| **ProductCategory_Lvl2** | Subcategory to which the product belongs |

# STORE (HACKATHON_DIMSTORE_SAN_VSHARED.CSV)

This dataset details the list of operational stores within the business unit including descriptive fields like region and whether they were online or offline

| Column | Description |
|---|---|
| **StoreKey** | Unique Store identifier |
| **DistributionChannel** | Distribution channel (e.g., Online or Physical) |
| **StoreType** | Store type (e.g., Type A, Type B, …) |
| **Region_Lvl1** | Region (e.g., Region A, Region B, Online) |
| **Region_Lvl2** | Region (more granular than Region_Lvl1) |

# TRANSACTION (HACKATHON_FACTSALESTRANSACTIONDATES_VSHARED.CSV)

- This dataset in an aggregation of raw underlying transaction data up to the day-store-product level. It contains summary information about the actual sales, units and discount and retail full price. It is also connected to additional information on the dates incl. weekdays and weekends
- It covers a time period from Jan 2020 to Dec 2022

| Column | Description |
|---|---|
| **TransactionDate** | Date of transaction |
| **DayOfWeek** | Weekday of the transaction date |
| **WeekendFlag** | Flag indicating whether the transaction took place during the weekend |
| **StoreKey** | Store identifier |
| **ProductKey** | Product identifier |
| **UnitVolume** | Units sold during transaction |
| **ActualSales** | Price paid by the customer |
| **SalesDiscount** | Discount for the customer (transaction level) |
| **RetailFullPrice** | Value of the transaction (without accounting for discounts) |

# TRANSACTION PROMOTION (HACKATHON_FACTSALESTRANSACTIONPROMOTION_VSHARED.CSV)

- This dataset in an aggregation of raw underlying transaction data up to the day-store-product-promotion level. It contains summary information about the discount given away per promotion
- It covers a time period from Jan 2020 to Dec 2022

| Column | Description |
|---|---|
| **TransactionDate** | Date of transaction |
| **StoreKey** | Store identifier |
| **ProductKey** | Product identifier |
| **PromotionKey** | Promotion identifier |

# CPI (CONSUMER PRICE INDEX_VSHARED.XLSX)

- This is an external dataset which measure the Consumer Price Index for Malaysia
- It covers a time period from Jan 2019 until Dec 2022

| Column | Description |
|---|---|
| **Date_monthly** | Reference date for monthly CPI |
| **CPI_monthly** | Monthly CPI level |
| **Date_daily** | Reference date for daily CPI |
| **CPI_daily** | Daily CPI level (monthly CPI transformed to daily granularity) |

# HOLIDAYS (HACKTHON_HOLIDAYSMY_VSHARED.CSV)

This is an external dataset which contains a list of public holidays in Malaysia

| Column | Description |
|---|---|
| **Holiday Description** | Name of public holidays in Malaysia |
| **Comments** | Any additional comments around nationwide participation |

# B.

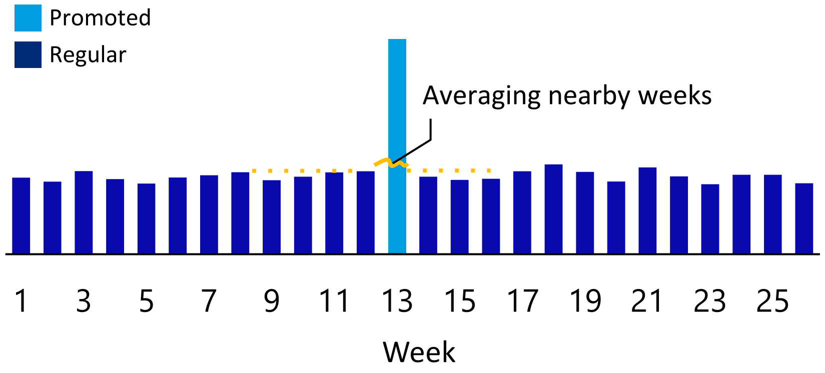
## INTRODUCTION TO MODELLING PRODUCT BASELINE AND UPLIFT

# BASELINE OF SALES CAN BE DETERMINED THROUGH A VARIETY OF METHODS

Illustrative approaches to modelling product baselines

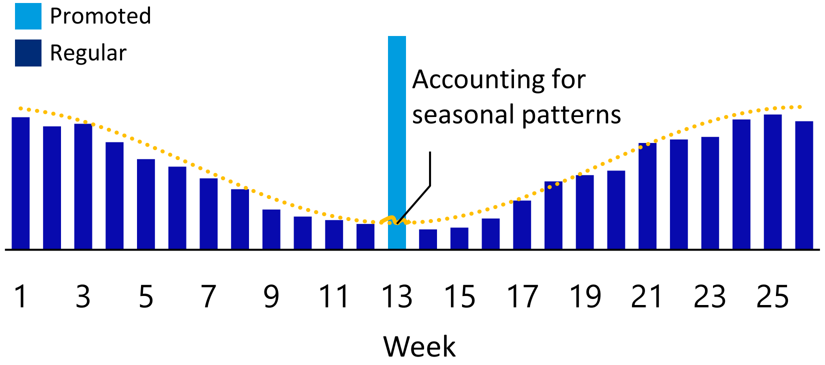
- The goal of modelling product baselines and uplift is to **understand what the sales for a product would have been if there had been no promotion**
- When looking at aggregated data across time, it is possible to “guess” what the baseline would be looking at the trends when looking at a simple example (1)
- As noise and complexity is added (4), additional factors need to be taken into account so a more complex approach must be followed

1



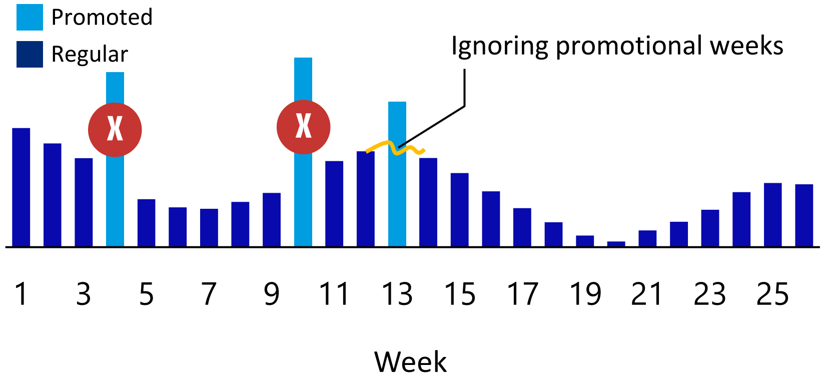
Averaging nearby weeks

2



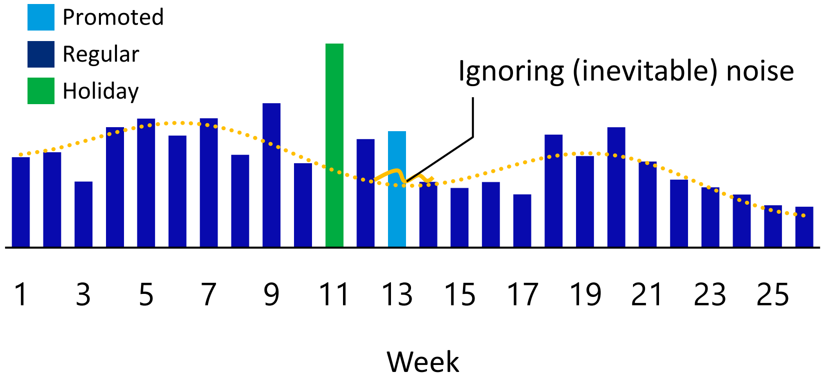
Accounting for seasonal patterns

3



Ignoring promotional weeks

4



Ignoring (inevitable) noise

# WE CAN TRAIN A MODEL TAKING MULTIPLE FACTORS INTO ACCOUNT TO PREDICT THE VOLUME IN THE BASELINE, AND THEN CONVERT THIS TO BACK INTO SALES

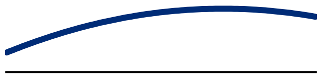
$$V(t) = \exp\left[\alpha + \underbrace{\beta_1 t + \beta_2 t^2}_{1} + \underbrace{\varepsilon_1 p_t}_{2} + \underbrace{\varepsilon_2 d_t}_{3} + \underbrace{\sum_{i=1}^{n} \lambda_i \mathrm{SEAS}_i(t)}_{4} + \underbrace{\sum_{j=1}^{m} \varphi_j \mathrm{HOL}_j(t)}_{5} + \cdots\right]$$

**Volume on product level**

**Constant**

**…is a non-linear function of**

### 1 Trends



Translates slow incremental trends in sales. Time in sequential weeks from start period

$\beta_1$: linear coefficient
$\beta_2$: quadratic coefficient

### 2 Promotions

Takes into account the impact on volumes of past promotions

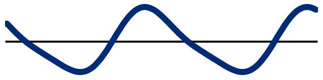
$\varepsilon_1$: promo flag coefficient
$p_{(t1)}$: promo flag

### 3 Promo discount rate

Reflects the impact of ongoing price changes on past volumes

$\varepsilon_2$: discount rate coefficient
$d_{t2}$: discount rate

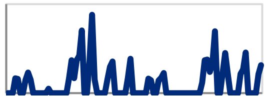
### 4 Seasonality



Terms of a Fourier series i.e. sin and cos terms with varying periods; reflect structural trends within one year

$\lambda_i$: Fourier terms

### 5 Holiday factors



Reflects the impact of holidays on volumes to be combined with holiday flags

$\varphi_j$: Holiday coefficient

**Additional factors may include**

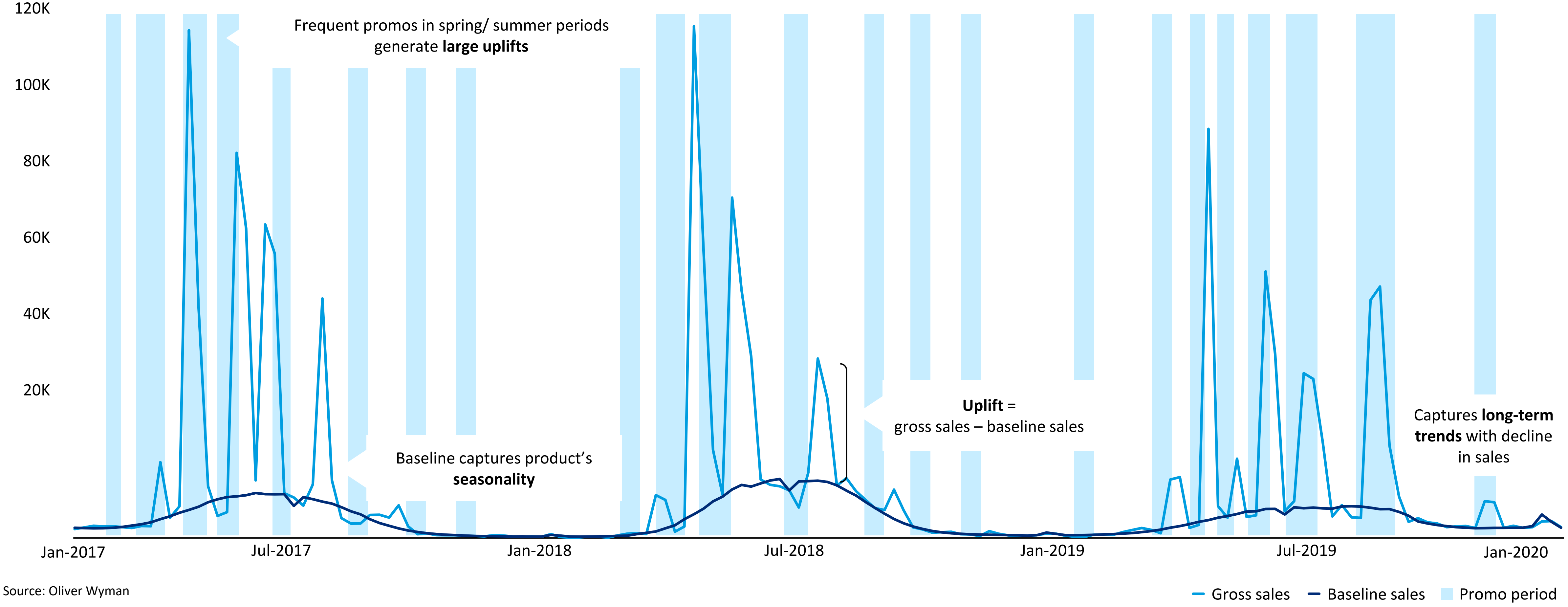
- Days of the week
- Weekend days
- Significant events outside of public holidays

*Please note, all factors may not be included in the final model if they are not significant*

# OUTPUT OF THE BASELINE MODEL SHOULD BE RESISTANT TO PROMOTIONAL PERIODS AND ADJUST FOR SEASONALITY

**Example**

Product gross sales and baseline sales (2017-2020, in €)



Frequent promos in spring/ summer periods generate **large uplifts**

Baseline captures product's **seasonality**

**Uplift** = gross sales – baseline sales

Captures **long-term trends** with decline in sales

Source: Oliver Wyman

OliverWyman

A business of Marsh McLennan