---

title: "Business Intelligence Report"

format: html

Introduction

```{r setup, echo=FALSE, error=FALSE, warning=FALSE, message=FALSE}

library(tidyverse)

library(gt)

knitr::opts_chunk$set(echo = FALSE,

warning = FALSE,

message = FALSE,

error = FALSE)

load("data/task1.RData")

```

I was asked us to evaluate the performance of a store trial which was performed in stores 77, 86, and 88. The first step is to aggregate purchases by customer and total sales at each store by month so that we can compare store performance by month and find a suitable control group to measure the trial group against.

```{r add_date_helpers}

store_data =

full_df %>%

filter(prod_brand!="Old El Paso") %>%
```

```
 mutate(year=year(date),
        month=month(date),
        day=day(date),
        wday=wday(date),
        .before = "date")

```

```{r monthly_summary}

store_monthly =
 store_data %>%
 summarize(
  across(.cols  = prod_qty:pack_size,
         .fns   = list(sum=sum),
         .names = "{col}"),
  transx = n(),
  .by = c(year,month,store_nbr,lylty_card_nbr)) %>%
 summarize(
  date = date(str_glue("{min(year)}-{min(month)}-{01}")),
  across(
   .cols  = prod_qty:transx,
   .fns   = list(sum=sum,
                 avg=mean),
   .names = "{col}_{fn}"),
  cust_nbr = n(),
```

```
   .by = c(year,month,store_nbr)) %>%
  arrange(year,month,store_nbr)

store_monthly %>%
  head() %>%
  gt()

```

Next we will visualize stores 77, 86, and 88 in order to get an idea of what their sales pattern looks like over the period data is available.

```{r linechart}

incomplete_stores =
  store_monthly %>%
  summarize(mnts=n(),.by = store_nbr) %>%
  filter(mnts<12) %>%
  pull(store_nbr)

treat_monthly =
  store_monthly %>%
  filter(!store_nbr %in% incomplete_stores) %>%
  mutate(trial_store =
         case_when(store_nbr %in% c(77, 86, 88) ~ 1,
                   TRUE ~ 0))
```

```
treat_monthly %>%
 ggplot() +
 geom_line(aes(x=date,
           y=tot_sales_sum,
           color=store_nbr),
       linewidth=1.2) +
 gghighlight::gghighlight(trial_store==1,
                  use_direct_label = FALSE) +
 scale_x_date(labels = scales::date_format(format = "%b"),
          breaks = scales::date_breaks()) +
 scale_y_continuous(labels = scales::dollar_format()) +
 labs(title = "Monthly chip sales by store",
    subtitle = "Highlighted lines represent trial group of stores 77, 86, and 88 all other stores are greyed out",
    y = "Chip sales",
    x = NULL,
    color = "Store number") +
 theme_minimal()
```

Based on the graph, store 77 seems to be a smaller store, store 86 seems to be mid-sized and store 88 is a large, top performing store. Control stores to compare to each trial store need to match the category of the trial store. Store 88 will likely have fewer candidates for matches since it is a bit of an outlier store.

In order to calculate similarity, we can first expand the store numbers to a grid of each store id paired with every other store id, then run a function to calculate the similarity between each pairing. This analysis uses pearson's correlation function and a normalized euclidean distance formula to calculate similarity. Each function makes up half the weight of a final scoring so that the top scoring store will be the most correlated and the closest size in terms of total sales during the pre-trial period. The table below shows the top 4 candidates for each trial store but only the top scoring store carries forward in further analysis. The graph shows the trial and control stores during the Pre-trial, Trial, and Post-trial periods

```{r sales_correl}

store_nbrs =
 treat_monthly %>%
 transmute(store_nbr = as.character(store_nbr)) %>%
 pull(store_nbr) %>%
 unique()

#Find each store's correlation with the trial stores to build a score to compare stores performance in terms of month to month changes. This combined with a store size comparison will be used to select control stores.
sales_correlations =
 expand_grid(c(77,86,88),store_nbrs) %>%
 select(nbr1=1,nbr2=2) %>%
 filter(nbr1!=nbr2) %>%
 mutate(across(everything(),as.numeric)) %>%
 mutate(correlations=
       map2_dbl(
        nbr1,
```

```
        nbr2,
        \(nbr1,nbr2){
         treat_monthly %>%
          select(year, month, store_nbr, tot_sales_sum) %>%
          #filter the monthly dataset to not include the treatment period, since we are interested in similarity between stores outside of the store layout trial period, but searching for differences during the layout trial. Pick two stores at a time and pivot wider in order to have two series to pass to correlations function.
          filter(!month %in% 2:6,
              store_nbr %in% c(nbr1, nbr2)) %>%
          pivot_wider(names_from = store_nbr,
                values_from = tot_sales_sum) %>%
          summarize(correlation = cor(x=pick(3),y=pick(4),
                            method = "pearson")) %>% pull()
        },.progress = TRUE))

```

```{r sales_magdistance}

#Find each store's sales "distance" (how different the size of sales is) from the trial stores to build a score to compare stores performance in terms of month to month size. This combined with a correlation comparison will be used to select control stores.

sales_distance =
 expand_grid(c(77,86,88),store_nbrs) %>%
 select(nbr1=1,nbr2=2) %>%
 filter(nbr1!=nbr2) %>%
```

```
 mutate(across(everything(),as.numeric)) %>%
 mutate(euclidean_dist=
        map2_dbl(
         nbr1,
         nbr2,
         \(nbr1,nbr2){
          treat_monthly %>%
            select(year, month, store_nbr, tot_sales_sum) %>%
            #filter the monthly dataset to not include the treatment period, since we are interested in similarity between stores outside of the store layout trial period, but searching for differences during the layout trial. Pick two stores at a time and pivot wider in order to have two series to pass to distance function.
            filter(!month %in% 2:6,
                   store_nbr %in% c(nbr1, nbr2)) %>%
            pivot_wider(names_from = store_nbr,
                        values_from = tot_sales_sum) %>%
            summarize(euclidean_dist = sqrt(sum((pick(3) - pick(4))^2))) %>% pull()
         },.progress = TRUE))


#Find the max and min euclidean distance for use in the magnitude difference score.
max_min_dist =
 sales_distance %>%
 summarize(
  across(
    .cols = euclidean_dist,
    .fns = list(max=max,
                min=min),
```

```
    .names = "{col}_{fn}"
  )
 )
```

```{r join&slice_heuristics}

comps = sales_correlations %>%
 inner_join(sales_distance,
        by = join_by(nbr1, nbr2)) %>%
 cross_join(max_min_dist) %>%
 #Create the store size similarity score and weight correlation and store size equally to create a score to select control stores.
 mutate(magnitude_distance =
       1 - ((euclidean_dist - euclidean_dist_min) /
            (euclidean_dist_max - euclidean_dist_min)),
      score = 0.5*correlations + 0.5*magnitude_distance) %>%
 select(nbr1:correlations, magnitude_distance:score) %>%
 group_by(nbr1) %>%
 slice_max(order_by = score,
        n = 4, with_ties = FALSE)
comps %>%
 gt(caption = "Top four stores similar to each trial store") %>%
 cols_label(nbr2="stores")
```

```{r filter_comps}
#grab lists of most similar stores to each trial store based on final scores.
comp_stores = comps %>%
 group_by(nbr1) %>%
 slice_max(order_by = score, n = 1) %>%
 pivot_longer(cols = nbr1:nbr2,
         values_to = "stores") %>%
 pull()



comp_df =
 treat_monthly %>%
 #filter original dataset for only trial and comp stores.
 filter(store_nbr %in% comp_stores)


comp_df %>%
 ggplot() +
 geom_rect(xmin=as_date("2018-06-01"),xmax=as_date("2019-01-15"),ymin=0,ymax=1500, fill="grey40", alpha=0.01) +
 geom_rect(xmin=as_date("2019-01-15"),xmax=as_date("2019-04-15"),ymin=0,ymax=1500, fill="firebrick1", alpha=0.01) +
 geom_rect(xmin=as_date("2019-04-15"),xmax=as_date("2019-07-01"),ymin=0,ymax=1500, fill="yellow3", alpha=0.01) +
 geom_text(x=as_date("2018-11-01"),
        y=500,
```

```
        label="Pre-trial period") +
 geom_text(x=as_date("2019-03-01"),
        y=500,
        label="Trial period") +
 geom_text(x=as_date("2019-05-15"),
        y=500,
        label="Post-trial\nperiod") +
 geom_line(aes(x=date,
           y=tot_sales_sum,
           color=store_nbr),
        linewidth=1.2) +
 scale_x_date(labels = scales::date_format(format = "%b"),
          breaks = scales::date_breaks()) +
 scale_y_continuous(labels = scales::dollar_format()) +
 labs(title = "Monthly chip sales by store",
    subtitle = "Comparison stores were picked based on similarity during the Pre-trial period",
    y = "Chip sales",
    x = NULL,
    color = "Store number") +
 theme_minimal()
```

```{r trial_stats}
```

```
trial_stats =
 comp_df %>%
 mutate(trial_period =
        case_when(
         date %in% as_date(paste0("2019-0",2:4,"-01")) ~ "trial",
         date %in% as_date(paste0("2019-0",5:6,"-01")) ~ "post_trial",
         TRUE~"pre_trial"
         )
       ) %>%
 summarize(
  across(
    .cols = prod_qty_sum:cust_nbr,
    .fns = list(avg=mean),
    .names = "{col}_{fn}"
  ),
  .by = c(store_nbr, trial_period)
) %>%
 select(store_nbr, trial_period,
       average_total_product_quantity=prod_qty_sum_avg,
       average_product_quantity_per_customer=prod_qty_avg_avg,
       average_total_sales=tot_sales_sum_avg,
       average_sales_per_customer=tot_sales_avg_avg,
       average_pack_size=pack_size_avg_avg,
       average_number_of_transactions=transx_sum_avg,
       average_number_of_transactions_per_customer=transx_avg_avg,
       average_number_of_customers=cust_nbr_avg) %>%
```

```
 pivot_longer(average_total_product_quantity:average_number_of_customers,
              names_to = "variable") %>%
 mutate(variable = str_replace_all(variable,"_"," "),
        variable = str_remove(variable,"average "))
```

The results table below shows store sales metrics from the pre-trial period before February 2019, the trial period from February to the end of April, and the post-trial period from May onward.

```{r results_table}
trial_results =
 trial_stats %>%
 mutate(store=
          case_when(
           store_nbr %in% c(77,86,88)
           ~ paste("Trial store",store_nbr),
           TRUE
           ~ paste("Control store",store_nbr))) %>%
 select(-store_nbr) %>%
 pivot_wider(names_from = trial_period) %>%
 mutate(delta = trial - pre_trial,
        pct_delta = delta/pre_trial,
        delta = scales::number(delta,
                               accuracy = .01,
```

```
                    style_positive = "plus",
                    style_negative = "minus"),
      pct_delta = scales::number(pct_delta,
                      accuracy = .01,
                      style_positive = "plus",
                      style_negative = "minus",
                      suffix = "%")
 )

trial_results %>%
 select(-post_trial) %>%
 gt(groupname_col = "store",
   caption = "Monthly store sales metrics for trial and control stores") %>%
 cols_label(.list =
          list(
           "variable"=" "
          ))

```

After a visual analysis of the results table it is clear that average monthly number of customers increases consistently in the trial groups and decreases consistently in the control groups. Since most other indicators stay mostly flat, increased customer numbers drives an increase in total sales in the trial stores while total sales fall in the control stores.

Further statistical analysis would be needed to determine the statistical significance of the increase in total sales.