Background information

You are part of Quantium retail analytics team and have been approachd by you client, the category managerfor chips,has asked us to test the impact of new trial layouts with data driven recommendation to whether or not the trial layout should be rolled out to all their stores.

The store trial layout was performed in store 77, 86 and 88.

1. Ask phase

The project aim is to test the impact of new trial layouts

2.Prepare phase

I am using python and Tabeau for this analysis, provided with QVI dataset.

In [1]:

```
#   import library

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import datetime
from scipy import stats
```

In [135]:

```
#    read dataset
qidf = pd.read_csv("QVI_data.csv")
```

In [136]:

```
#    few first rows
qidf.head()
```

Out[136]:

	LYLTY_CARD_NBR	DATE	STORE_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES	PACK_SIZE	BRAND	LIFESTAGE	PRE
0	1000	2018-10-17	1	1	5	Natural Chip Compny SeaSalt175g	2	6.0	175	NATURAL	YOUNG SINGLES/COUPLES	
1	1002	2018-09-16	1	2	58	Red Rock Deli Chikn&Garlic Aioli 150g	1	2.7	150	RRD	YOUNG SINGLES/COUPLES	
2	1003	2019-03-07	1	3	52	Grain Waves Sour Cream&Chives 210G	1	3.6	210	GRNWVES	YOUNG FAMILIES	
3	1003	2019-03-08	1	4	106	Natural ChipCo Hony Soy Chckn175g	1	3.0	175	NATURAL	YOUNG FAMILIES	
4	1004	2018-11-02	1	5	96	WW Original Stacked Chips 160g	1	1.9	160	WOOLWORTHS	OLDER SINGLES/COUPLES	

In [137]:

```
qidf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 264834 entries, 0 to 264833
Data columns (total 12 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   LYLTY_CARD_NBR    264834 non-null  int64  
 1   DATE              264834 non-null  object 
 2   STORE_NBR         264834 non-null  int64  
 3   TXN_ID            264834 non-null  int64  
 4   PROD_NBR          264834 non-null  int64  
 5   PROD_NAME         264834 non-null  object 
 6   PROD_QTY          264834 non-null  int64  
 7   TOT_SALES         264834 non-null  float64
 8   PACK_SIZE         264834 non-null  int64  
 9   BRAND             264834 non-null  object 
 10  LIFESTAGE         264834 non-null  object 
 11  PREMIUM_CUSTOMER  264834 non-null  object 
dtypes: float64(1), int64(6), object(5)
memory usage: 24.2+ MB
```

No missing value

```
In [138]: qidf.duplicated().sum() # check for duplicate
```

```
Out[138]: 1
```

```
In [151]: qidf = qidf.drop_duplicates() # remove duplicate
```

```
In [153]: qidf.duplicated().sum() # validate
```

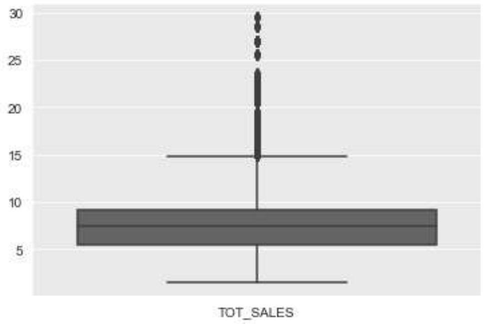
```
Out[153]: 0
```

```
In [140]: qidf.dtypes
```

```
Out[140]: LYLTY_CARD_NBR        int64
          DATE                 object
          STORE_NBR             int64
          TXN_ID                int64
          PROD_NBR              int64
          PROD_NAME            object
          PROD_QTY              int64
          TOT_SALES           float64
          PACK_SIZE             int64
          BRAND                object
          LIFESTAGE            object
          PREMIUM_CUSTOMER     object
          dtype: object
```

Check for outliers

```
In [141]: sns.boxplot(data = qidf[["TOT_SALES"]], showfliers = True);
```



```
In [144]: qidf["Z_SCORES"] = stats.zscore(qidf["TOT_SALES"]) #   z_score column
```

```
In [145]: qidf.head(3)
```

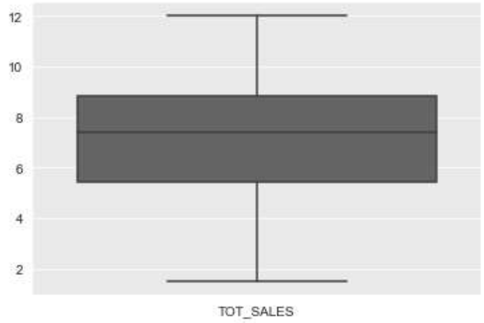
Out[145]:

ATE	STORE_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES	PACK_SIZE	BRAND	LIFESTAGE	PREMIUM_CUSTOMER	Z_SCORES
18- -17	1	1	5	Natural Chip Compny SeaSalt175g	2	6.0	175	NATURAL	YOUNG SINGLES/COUPLES	Premium	-0.514137
18- -16	1	2	58	Red Rock Deli Chikn&Garlic Aioli 150g	1	2.7	150	RRD	YOUNG SINGLES/COUPLES	Mainstream	-1.819912
19- -07	1	3	52	Grain Waves Sour Cream&Chives 210G	1	3.6	210	GRNWVES	YOUNG FAMILIES	Budget	-1.463791

```
In [148]: qidf = qidf[qidf["Z_SCORES"] < 2]
```

In [149]:

```
sns.boxplot(data = qidf[["TOT_SALES"]], showfliers = True); #    validate
```



In [154]:

```
df0 = qidf
```

4. Analyze phase

In [155]:

```
df0['DATE'] = pd.to_datetime(df0["DATE"], format="%Y/%m/%d") #    format to datetime datatype
df0["SALES_PRICE"] = df0.TOT_SALES/df0.PROD_QTY
df0["SALES_REVENUE"] = df0.PROD_QTY * df0.SALES_PRICE
```

In [156]:

```
def group(number):
    ''' A program written to group stores by trial layout, stores 77, 86 and 88 are labelled "treatment_num", while others a
    if number == 77:
        store = "treatment_num"
    elif number == 86:
        store = "treatment_num"
    elif number == 88:
        store = "treatment_num"
    else:
        store = "controlled_num"
    return store

```

In [157]:

```

df0["STORE_CAT"] = df0.STORE_NBR.apply(group) # apply function
```

In [158]:

```
df0["REVENUE_PERCENTAGE"] = (df0.SALES_REVENUE/df0.SALES_REVENUE.sum())* 100  #  sales revenue in percentage
```

In [159]:

```
df0.head()  # first few rows
```

Out[159]:

PACK_SIZE	BRAND	LIFESTAGE	PREMIUM_CUSTOMER	Z_SCORES	SALES_PRICE	SALES_REVENUE	STORE_CAT	REVENUE_PERCENTAGE
175	NATURAL	YOUNG SINGLES/COUPLES	Premium	-0.514137	3.0	6.0	controlled_num	0.000319
150	RRD	YOUNG SINGLES/COUPLES	Mainstream	-1.819912	2.7	2.7	controlled_num	0.000144
210	GRNWVES	YOUNG FAMILIES	Budget	-1.463791	3.6	3.6	controlled_num	0.000191
175	NATURAL	YOUNG FAMILIES	Budget	-1.701205	3.0	3.0	controlled_num	0.000159
160	WOOLWORTHS	OLDER SINGLES/COUPLES	Mainstream	-2.136463	1.9	1.9	controlled_num	0.000101

In [160]:

```
df0["STORE_CAT"].unique()  #  validate
```

Out[160]:

```
array(['controlled_num', 'treatment_num'], dtype=object)
```

In [161]:
```
#  subset by store category

controlled_df = df0[df0["STORE_CAT"] == 'controlled_num']
treatment_df = df0[df0["STORE_CAT"] == 'treatment_num']
```

In [162]:
```
controlled_df.shape
```

Out[162]:
```
(257228, 17)
```

In [163]:
```
treatment_df.shape
```

Out[163]:
```
(3922, 17)
```

In [164]:
```
treatment_df.size
```

Out[164]:
```
66674
```

In [165]:
```
#  sampling 20 records by replacement

controlled_sample_df = controlled_df.sample(n = 20, random_state = 42, replace = True)
treatment_sample_df = treatment_df.sample(n = 20, random_state = 41, replace = True)
```

In [166]:
```
controlled_sample_df.shape
```

Out[166]:
```
(20, 17)
```

In [167]:
```
treatment_sample_df.shape
```

Out[167]:
```
(20, 17)
```

In [168]:
```
# mean of store categories

controlled_sample_mean = controlled_sample_df["TOT_SALES"].mean()
treatment_sample_mean = treatment_sample_df["TOT_SALES"].mean()
```

In [169]:
```
controlled_sample_mean
```

Out[169]:
```
7.825000000000001
```

In [170]:
```
treatment_sample_mean
```

Out[170]:
```
7.635000000000001
```

In [171]:
```
difference = treatment_sample_mean - controlled_sample_mean
```

In [172]:
```
difference  # difference in mean
```

Out[172]:
```
-0.1900000000000004
```

5.Construct phase

Coduct hypothesis test to see if the observed difference is statistically significant or due to chance

Hypothesis by total sales

Null hypothesis: There is no difference between the means of the two groups by sales

Alternative hypothesis: There is a difference between the means of the two groups by sales

In [173]:
```
# For this analysis, the significance level is 5%
significance_level = 0.05
significance_level
```

Out[173]:
```
0.05
```

In [174]:
```
stats.ttest_ind(a = controlled_sample_df["TOT_SALES"], b = treatment_sample_df["TOT_SALES"], equal_var = False)
```

Out[174]:
```
Ttest_indResult(statistic=0.28986882695127353, pvalue=0.7734934129672759)
```

Hypothesis by revenue

Null hypothesis: There is no difference between the means of the two groups by revenue

Alternative hypothesis: There is a difference between the means of the two groups by revenue

In [175]:

```
# For this analysis, the significance level is 5%
significance_level = 0.05
significance_level
```

Out[175]:

```
0.05
```

In [176]:

```
stats.ttest_ind(a = controlled_sample_df["SALES_REVENUE"], b = treatment_sample_df["SALES_REVENUE"], equal_var = False)
```

Out[176]:

```
Ttest_indResult(statistic=0.28986882695127353, pvalue=0.7734934129672759)
```

Since the p values is way greater than the significant lever we fail to reject the null hypothesis

6. Share phase

In [177]:

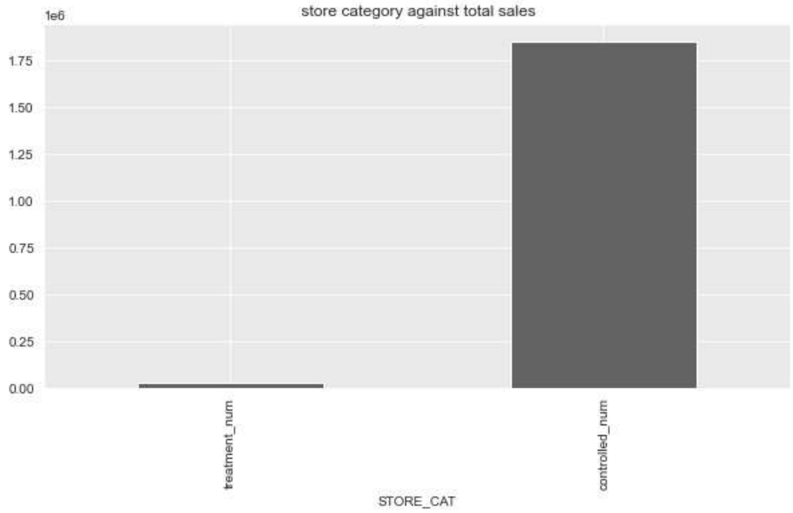
```
df0.groupby('STORE_CAT').TOT_SALES.sum().sort_values()
```

Out[177]:

```
STORE_CAT
treatment_num       29276.00
controlled_num    1851920.55
Name: TOT_SALES, dtype: float64
```

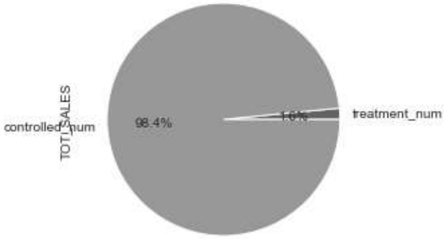
In [178]:

```
df0.groupby('STORE_CAT').TOT_SALES.sum().sort_values().plot(kind='bar',figsize=(10,5));
sns.set_style("darkgrid")
plt.title(" store category against total sales");
```



In [185]:

```
plt.figure(figsize=(4,4))
df.groupby("STORE_CAT").TOT_SALES.sum().sort_values().plot.pie(autopct="%1.1f%%")
plt.show()
```



In [179]:

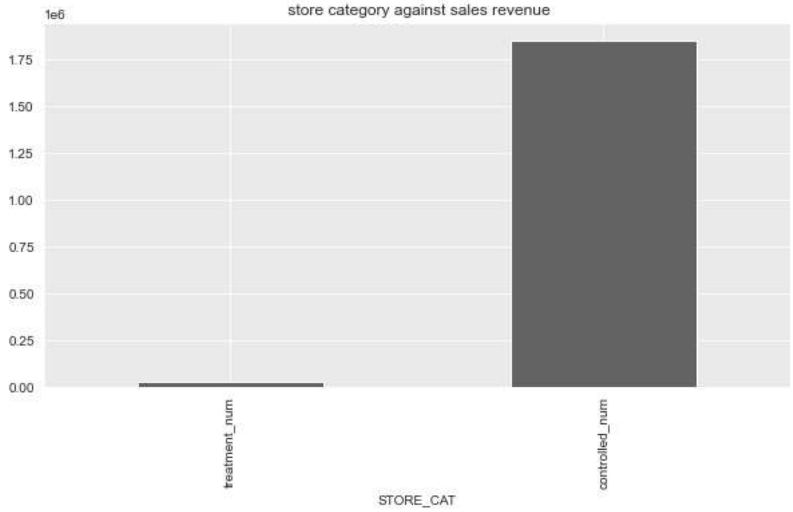
```
df0.groupby('STORE_CAT').SALES_REVENUE.sum().sort_values()
```

Out[179]:

```
STORE_CAT
treatment_num       29276.00
controlled_num    1851920.55
Name: SALES_REVENUE, dtype: float64
```

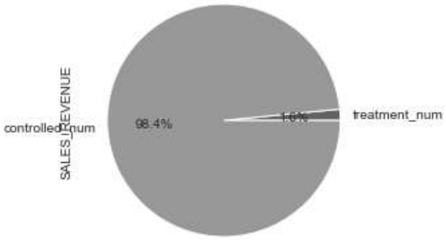
In [186]:

```
df0.groupby('STORE_CAT').SALES_REVENUE.sum().sort_values().plot(kind='bar',figsize=(10,5));
sns.set_style("darkgrid")
plt.title(" store category against sales revenue");
```



In [187]:

```
plt.figure(figsize=(4,4))
df.groupby("STORE_CAT").SALES_REVENUE.sum().sort_values().plot.pie(autopct="%1.1f%%")
plt.show()
```



In [181]:

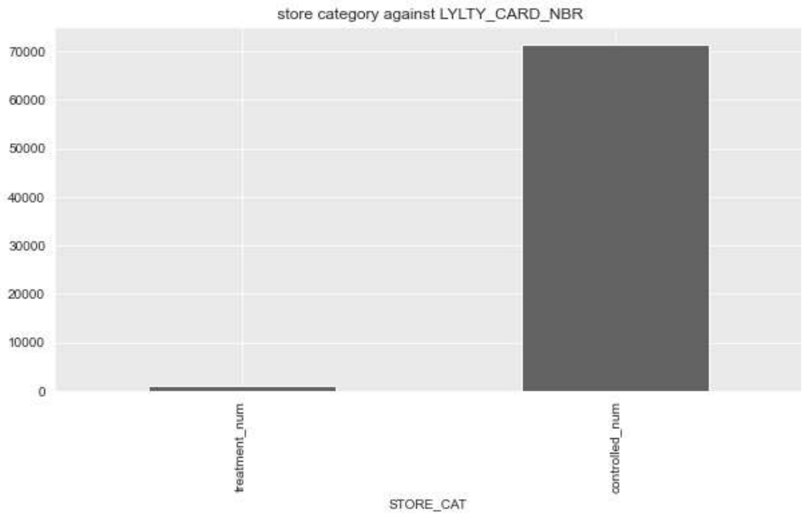
```
df0.groupby('STORE_CAT').LYLTY_CARD_NBR.nunique().sort_values()
```

Out[181]:

```
STORE_CAT
treatment_num      1014
controlled_num    71405
Name: LYLTY_CARD_NBR, dtype: int64
```

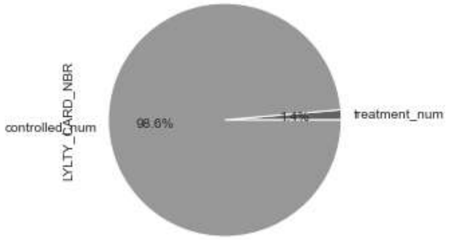
In [182]:

```
df0.groupby('STORE_CAT').LYLTY_CARD_NBR.nunique().sort_values().plot(kind='bar',figsize=(10,5));
sns.set_style("darkgrid")
plt.title(" store category against LYLTY_CARD_NBR");
```



In [188]:

```
plt.figure(figsize=(4,4))
df.groupby("STORE_CAT").LYLTY_CARD_NBR.nunique().sort_values().plot.pie(autopct="%1.1f%%")
plt.show()
```



In [183]:

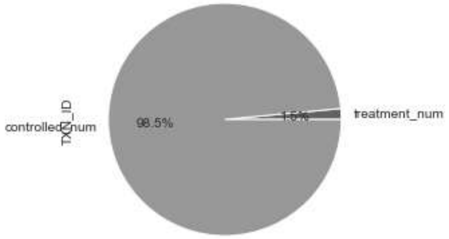
```
df.groupby("STORE_CAT").TXN_ID.nunique().sort_values()
```

Out[183]:

```
STORE_CAT
treatment_num       3894
controlled_num    255591
Name: TXN_ID, dtype: int64
```

In [184]:

```
plt.figure(figsize=(4,4))
df.groupby("STORE_CAT").TXN_ID.nunique().sort_values().plot.pie(autopct="%1.1f%%")
plt.show()
```



7. Act phase

Key findings

The stores the trial layout was performed in contributed:

1. 1.6 % of the total sales
2. 1.4 % of the total customers

3. 1.6 % of the total transactions

Recommendation

The new trial layout should not be rolled out to other stores

In [ ]:

```
```