A quick glance at the data

The data is from a bakery company, including inventory demand of over 1 million stores on more than 50 thousands bakery products for the past 7 weeks. The goal is to develop a model to accurately forecast inventory demand based on historical sales data. I used Python to analyze the data. The purpose of this pdf file is to have a quick glance at the historical sales data to have a general sense of how the data looks like.

First, a quick scan of the data size, ~ 74 million rows of data

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import math
import matplotlib.pyplot as plt
import os
from IPython.display import display_markdown as mkdown # as print

train = pd.read_csv('../Data/train.csv', usecols=['Demanda_uni_equil'])
print ("train",len(train))
```

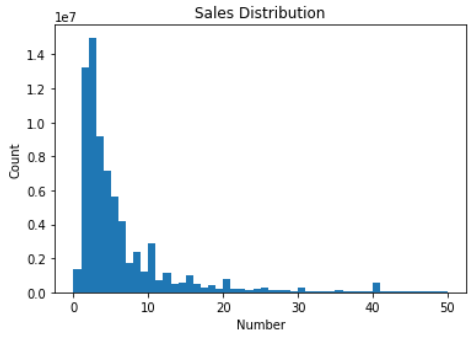
```
train 74180464
```

Then we can generate 1D and 2D histogram to show the average number for each product sold at the store in a week. From the histogram, it is easy to conclude that the average number sold for each product at a local store is around 3. This number is important when a new product is introduced to the market.

```
sales = train['Demanda_uni_equil'].tolist()

def label_plot(title, x, y):
    plt.title(title)
    plt.xlabel(x)
    plt.ylabel(y)

plt.hist(sales, bins=50, range=(0, 50))
label_plot('Sales Distribution', 'Number', 'Count')
plt.show()
```



```
itemID = train.loc[train.Demanda_uni_equil < 20].index.tolist()
sales = train.loc[train.Demanda_uni_equil < 20].Demanda_uni_equil.tolist()

plt.hist2d(itemID, sales, bins=[50,20])
label_plot('Histogram of the number sold for each product', 'Index', 'Number')
plt.show()
```