YEDİTEPE UNIVERSITY

CSE 464
INTRODUCTION TO DATA SCIENCE & BIG DATA ANALYTICS

CASE STUDY 1

Nilhan Süer
Computer Science Engineering
20190702121

Table of Content

1. Short Story of Business/Organization Challenge

Our company that is a UK-based and registered non-store online retail sells unique all-occasion gifts. Over the past year, we have faced several challanges in maintaining our competitive edge and meeting our sales targets. We get feedback that we had problems to design new products and improve features of the products with respect to the customer needs.

Our company's sales figures and company Key Performance Indicators decreased and we couldn't achieve our target rates. So that, the factor of reduced sales make our concerns increase about our business strategy and we began to think that we didn't make the right decisions.

2. Problem Summary/Definition

Our first problem is the product innovation and customer satisfaction. There is a lack of connectivity between the products offered by the company and the needs of the customer base. So, this results in reduced sales due to dissatisfaction. As we see in the dataset analysis, various items are listed with their descriptions and quantities sold. But there are some missing factors on the dataset such as customer satisfaction or product performance, it shows they only focus on the transactional data.

The second problem is that, the sales performance and the Key Performance Indicators are below targeted rates which indicates the lack of efficiency in making decisions and business strategy. Although the transactions and revenue data are recorded in the dataset, the quantities and unit prices of items sold shows that the company's sales performance is below the average rate. In addition, some detailed analyses like the revenue growth are missing, so that we can understand that the strategic insights are not enough to improve the sales performances.

The third problem is the competitors and market differentiation. Some of the competitors have imitated our product ideas. So that, a threat to our market position occured. In out dataset, there are various informations about customer Ids and countries, but there is not enough details about customer segmentation, market trends or the activities that these competitors done. If we don't do detailed analyses of customer tendency, we may face with some problems to maintain out competitive advantage.

As a result, our company struggles with the problems in product innovation, sales performance and the market differentiation, that indicates our insufficiency of detailed data analysis and taking strategic decisions.

3. Solution/ Recommendations/Decisions

Firstly, as the data scientist group; we begin with loading the data. Then, we understand the data set by take a look at their mean, min, count, max and std values to decide which features we will use. We also clean data by converting null values.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_excel('OnlineRetail.xlsx') # 1st step

df.dtypes # 2nd step
```

```
InvoiceNo              object
StockCode              object
Description            object
Quantity                int64
InvoiceDate    datetime64[ns]
UnitPrice             float64
CustomerID            float64
Country                object
dtype: object
```

```
df.describe() # 2nd step
```

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	541909.000000	541909	541909.000000	406829.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114	15287.690570
min	-80995.000000	2010-12-01 08:26:00	-11062.060000	12346.000000
25%	1.000000	2011-03-28 11:34:00	1.250000	13953.000000
50%	3.000000	2011-07-19 17:17:00	2.080000	15152.000000
75%	10.000000	2011-10-19 11:27:00	4.130000	16791.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	218.081158	NaN	96.759853	1713.600303

By analysing the data features, we decide on which columns to use and selected the following features and get rid of 'InvonceNo' column due to its uselessness.

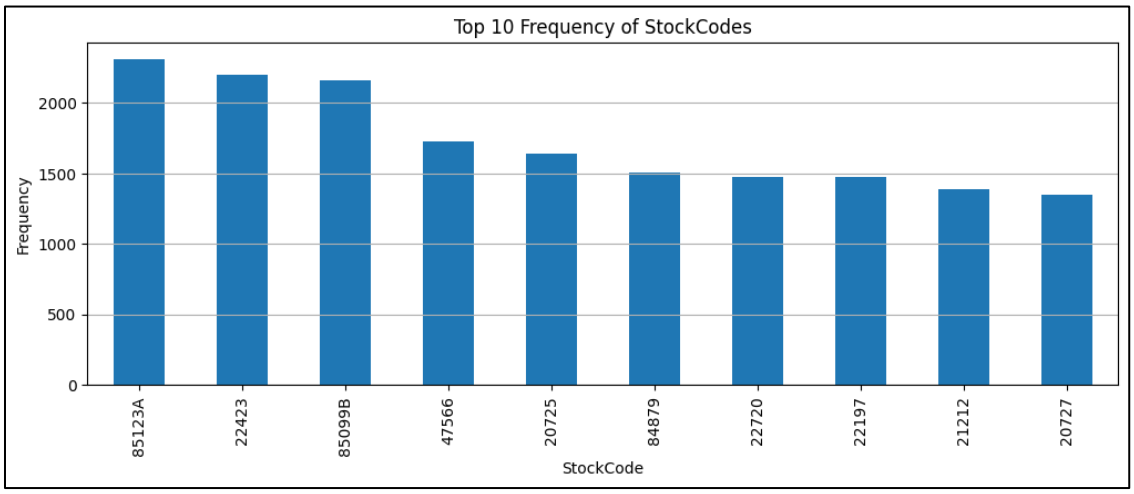
```
# 4th step - Get rid of 'InvonceNo' column
selected_columns = ['StockCode', 'Description', 'Quantity', 'InvoiceDate', 'UnitPrice', 'CustomerID', 'Country']
df_selected = df[selected_columns]

df_selected.head() # 5th step
```

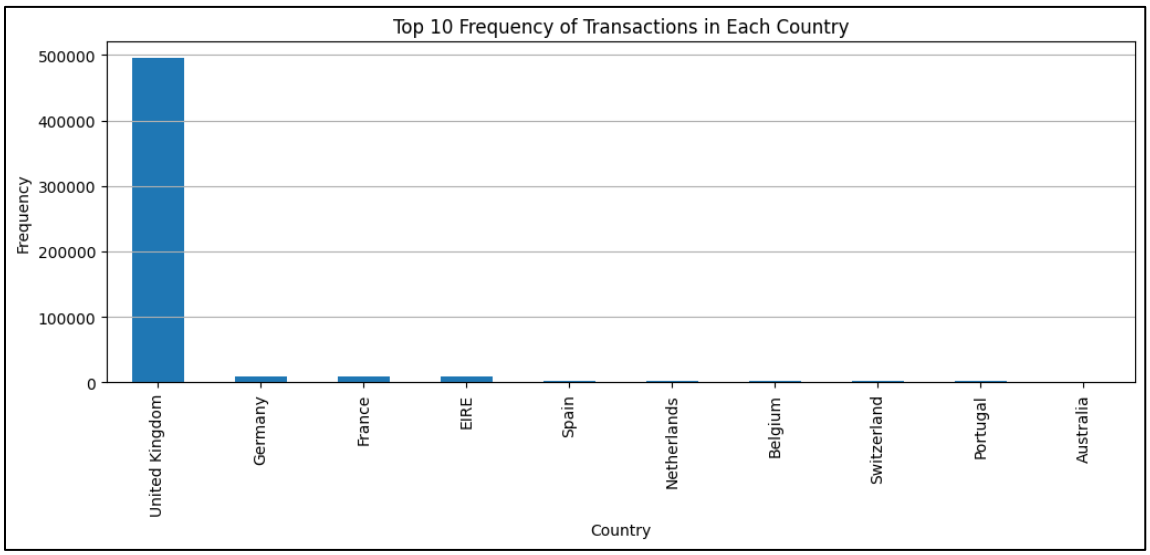
	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

In the analyzing part, we use descriptive analytics methods to understand what happened in the past. And we prefer 4 supporting methods such as measures of frequency, measures of central tendency, measures of dispersion and measures of position.

3.1. Measures of Frequency

By visualizing the frequency of each StockCode which represents the items, we try to understand which products or countries are more prevalent in the data set.





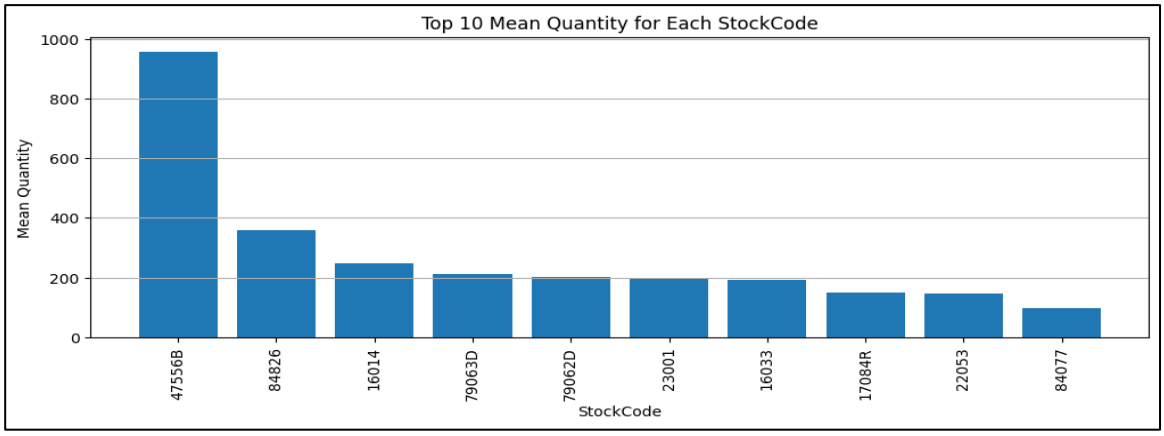

```
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

df['Hour'] = df['InvoiceDate'].dt.hour

# Calculate the frequency of transactions for each hour of the day
hourly_frequency = df['Hour'].value_counts().sort_index()

plt.figure(figsize=(12, 6))
hourly_frequency.plot(kind='bar', color='skyblue')
plt.title('Frequency of Transactions by Hour of the Day')
plt.xlabel('Hour of the Day')
plt.ylabel('Frequency')
plt.xticks(rotation=0)
plt.grid(axis='y')
plt.show()
```

To understand which times of the day the sales are at their highest, we can use measures of frequency such as counts or frequencies of transactions at different times of the day. We can calculate the frequency of transactions for each hour of the day and plot it to visualize the busiest times.

3.2. Measures of Central Tendency

For each StockCode in our data set, we can calculate the mean value for each StockCode. To achieve the result, we plot the StockCode versus their mean quantities. Thanks to this method, we can understand the central tendency of the 'Quantity' variable for each item.

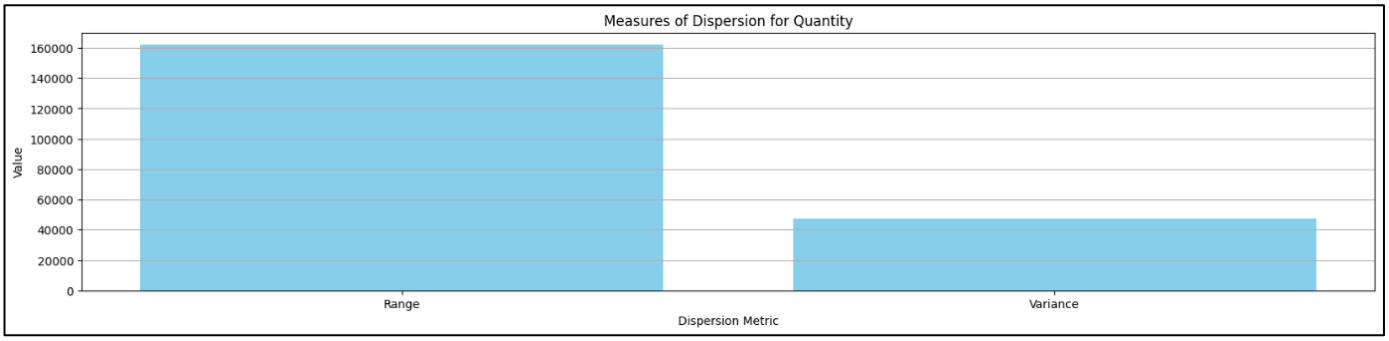
```
# Calculate the mean value for each StockCode
mean_values = df.groupby('StockCode')['Quantity'].mean().reset_index()

mean_values_sorted = mean_values.sort_values(by='Quantity', ascending=False)

top_10_mean_values = mean_values_sorted.head(10)

# StockCode versus their means
plt.figure(figsize=(12, 4))
plt.bar(top_10_mean_values['StockCode'].astype(str), top_10_mean_values['Quantity'])
plt.title('Top 10 Mean Quantity for Each StockCode')
plt.xlabel('StockCode')
plt.ylabel('Mean Quantity')
plt.xticks(rotation=90)
plt.grid(axis='y')
plt.show()
```



3.3. Measures of Dispersion – Range / Variance

Using 'Measures of Dispersion' method, we can get the range and variance values to be informed about how much our item's quantities vary and what their min and max values are. When we examine the graph, we can see that the range of the quantity values are at higher values. So maybe, it can be better to reduce that variety between the product quantities, so that our sales can be increased.

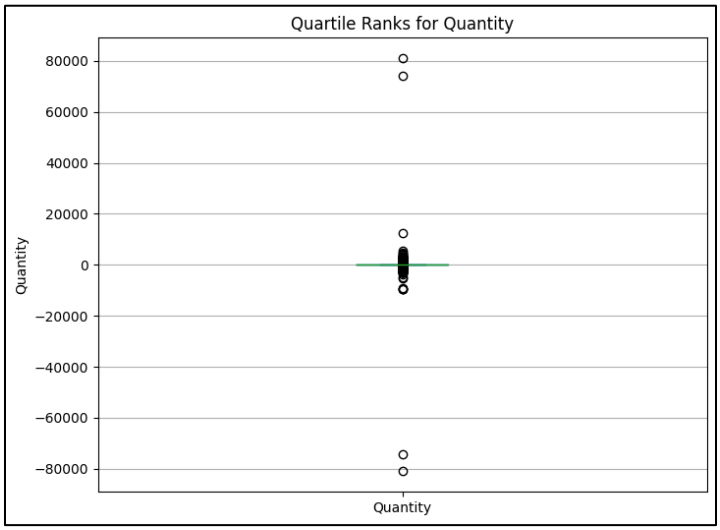
```
# Calculate range, variance for Quantity
quantity_range = df['Quantity'].max() - df['Quantity'].min()
quantity_variance = df['Quantity'].var()

dispersion_metrics = pd.DataFrame({
    'Metric': ['Range', 'Variance'],
    'Value': [quantity_range, quantity_variance]
})

plt.figure(figsize=(20, 4))
plt.bar(dispersion_metrics['Metric'], dispersion_metrics['Value'], color='skyblue')
plt.title('Measures of Dispersion for Quantity')
plt.xlabel('Dispersion Metric')
plt.ylabel('Value')
plt.grid(axis='y')
plt.show()
```



3.4. Measures of Dispersion – Quartiles and Box Plots

By the 'Measures of Position' method, we can be informed about quartiles, median and potential outliers by creating box plots. As a result, Q1 is 1.0, which means that 25% of the data points in the dataset have a value of 1.0 or lower; the median is 3.0, indicating that half of the data points in the dataset have a value of 3.0 or lower, and the other half have a value higher than 3.0; Q3 is 10.0, meaning that 75% of the data points in the dataset have a value of 10.0 or lower.

```
# Calculate quartile ranks for Quantity
quartiles = df['Quantity'].quantile([0.25, 0.5, 0.75])
q1, median, q3 = quartiles

plt.figure(figsize=(8, 6))
df['Quantity'].plot(kind='box')
plt.title('Quartile Ranks for Quantity')
plt.ylabel('Quantity')
plt.grid(axis='y')
plt.show()

print("First Quartile (Q1):", q1)
print("Median (Q2):", median)
print("Third Quartile (Q3):", q3)
```



4. Follow Up & Evaluation Plan

We have faced with several problems in our online retail sales company such as decreased sales performance, dissatisfaction of customers about products and the competitors factor. To come up with a solution, we used descriptive methods to analyze our data set that we obtain from customer tendencies in the past.

In summary, by creating some strategies that I mentioned above, we can make our products match our customer base needs and achieve highest sales rates. In addition, by increasing the frequency of our sales intervals and taking quick decisions, we can compete successfully with other companies. Finally, we can improve our business strategies according to current datas by making data analysis at regular intervals.

5. References

- CSE 464 Course Classroom Notes

- https://archive.ics.uci.edu/dataset/352/online+retail

- https://www.knowledgehut.com/blog/data-science/descriptive-analytics