The following document will show how to take a data base of a company's business that consists of food products which are often purchased as gifts during the Christmas season. It will show how load the database into PgAdmin, do exploratory data analysis, calculate RFM, and find the customer lifetime value.

1. Creating and loading a database in Postgresql

To manage the creation of a database the software PgAdmin 4 v4 was used as a managing tool for PostgreSQL the data. The data set provided comes from a MultiChannel Gift Company and it includes four files, the first field in all 4 files is a customer ID which will be used to link the files together;

The first step is the installation of both PosgreSQL and PgAdmin. After the installation of the software we begin by creating a local server within PgAdmin. For the purpose of this assignment the server was named "my_local_server" and it was configured with user "posgres" and default port "5432".

We start by creating the tables to load the necessary files. For three out of the four files, this was a straightforward step. The following template was sufficient for creating each table for the files of contacts, lines, and orders.

```
CREATE TABLE table_name (
    column1 datatype,
    column2 datatype,
    column3 datatype,
    ...
);
```

After each table was was created we can utilize the inbuilt tools of $PgAdmin > Tools > Import/Export > *.csv$, for quickly loading the data. It is important in this step that column names match the files and specify that the data contains headers. Additionally, after loading we rename the columns to have consistent names that follow camel case convention.

However, the summary file required quite a bit of data cleaning before it could be loaded into PgAdmin. The primary reason was that the purchase behavior variable was spread over several years, categories, and season. For this analysis we will focus on these columns since the later part of the files only included categorical information.

For the data wrangling process, a jupyter notebook with python code and the pandas module was used to melt together the relevant columns. Afterwards, each individual variable from the purchase behavior was extracted so the file could be imported into the database. The code to process the data is attached as a separate file with this document. The variables that we retained for analysis are the following:

```
CREATE TABLE customers (
    cust_id integer,
    scf_code text,
    purchase integer,
    channel text,
    season text,
    year integer,
    measure text
);
```

There are two important consideration that were made when cleaning this file. One, it was decided to retain the values of the period prior to 2004, these were marked as "Pre04" in the excel files and were stored under a special season 'P'. Two, there was a column containing information about a "Recency measure", these columns were dropped in order to keep a file format that would be consistent with tidy data. Finally we demonstrate how the database should look on the platform:

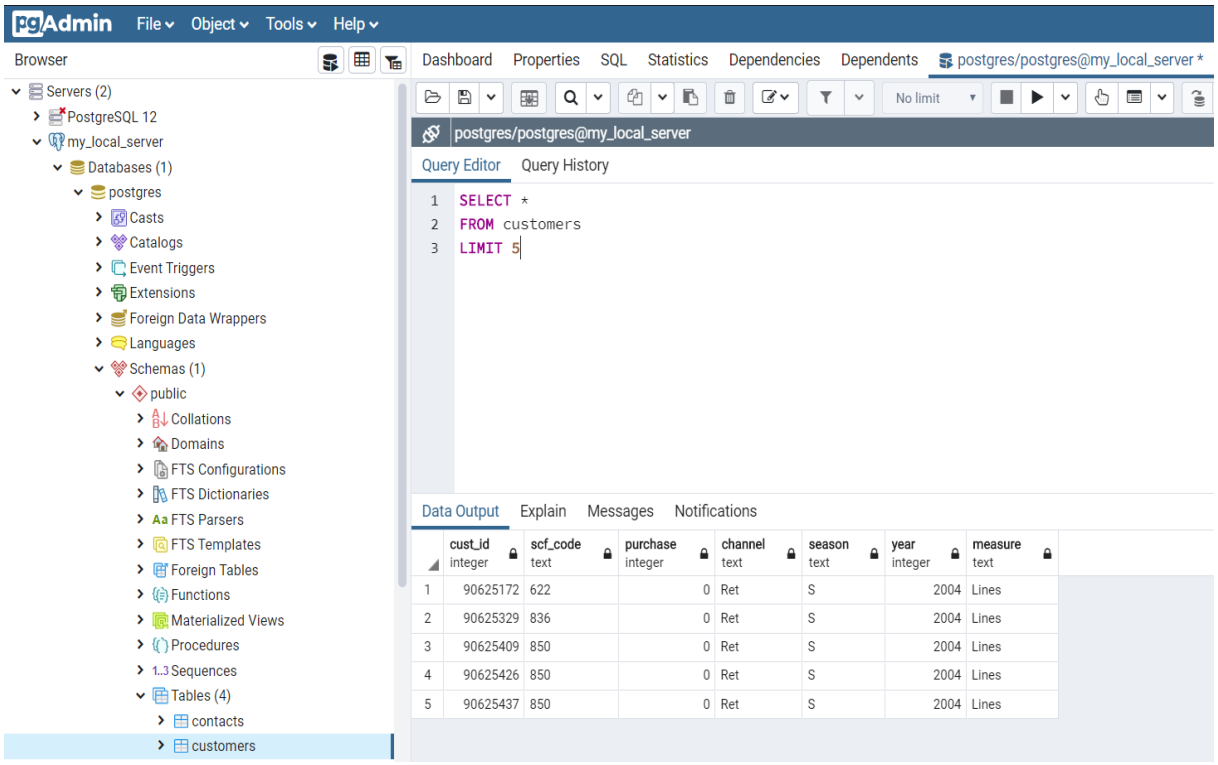

Figure 1: Image capture of pgAdmin with database loaded

1.1 Query to identify the top 5 customer locations by average spend.

The following sql query identifies the top 5 customer locations by average spend, this calculation was extracted by taking the zip codes from the purchasing behaviour of the customers to determine the geographical locations that on average generated the most amount of sales. The auxiliary code for loading and cleaning the data is the attached as a jupyter notebook html file, additionally the code for the SQL queries is also included as a separate file. The request to provide the SQL query is shown in the following code:

```
SELECT scf_code, ROUND(AVG(purchase),2) AS average_spend
FROM customers
WHERE scf_code IS NOT NULL
AND purchase > 0
GROUP BY scf_code
ORDER BY average_spend DESC
LIMIT 5
```

2. Exploratory Data Analysis

This section consists of diving into the data to gain "insights" into the effectiveness of the various direct-marketing channels. Specifically, the point of interested was to focus on the relation between catalog mailing vs email for the marketing efforts.

Before we begin the exploration of the data for that specific emphasis, it is crucial to get an overall picture of the company in the last couple of years. Our first graph shows the sum of total purchases by different channels. We can see that there was a big increase in the number of internet purchases from 2004 to 2005 and it has kept growing. The last year there was a big drop in total purchases for 'Cat', and Retail has stagnated in sales over the same period of 2006-2007.

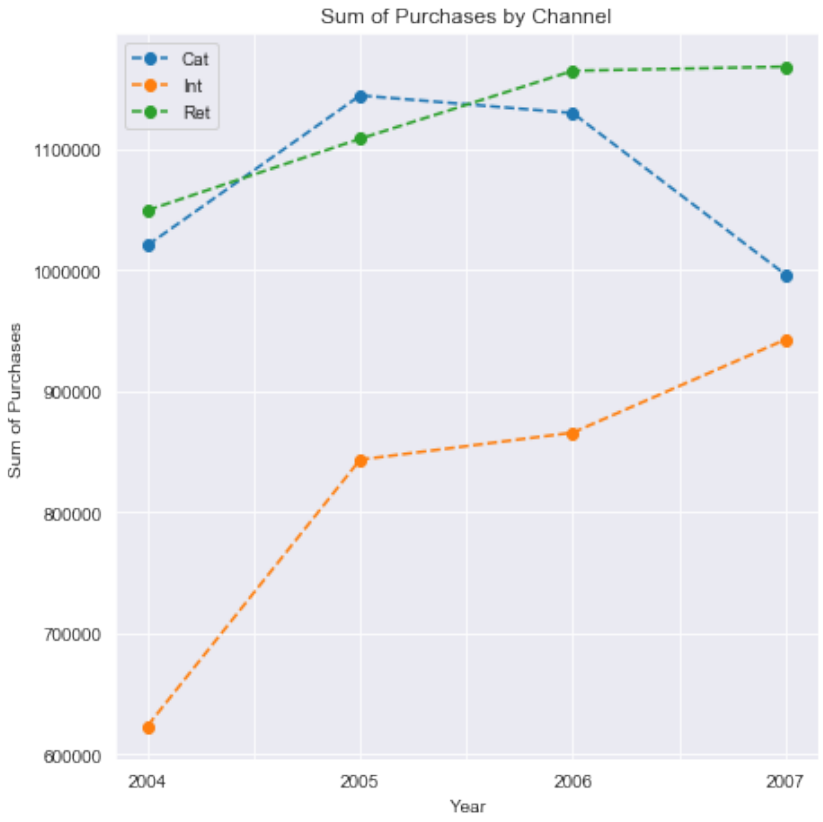

Figure 2: Sales across different Channels for the period 2004-2007

2.a Data-analytical Question 1: Which contact type is more effective per month?

Moving unto the comparison of catalog vs email. There is a crucial question about demonstrating which channel has a better effect in increasing sales and we will focus on the impact that each channel has across months. The vast majority of the communications through the marketing department are sent out as emails. This is likely due to the lower price of sending email marketing than catalog marketing. The distributions of both types of channels across the different month of the year can be seen as follows:

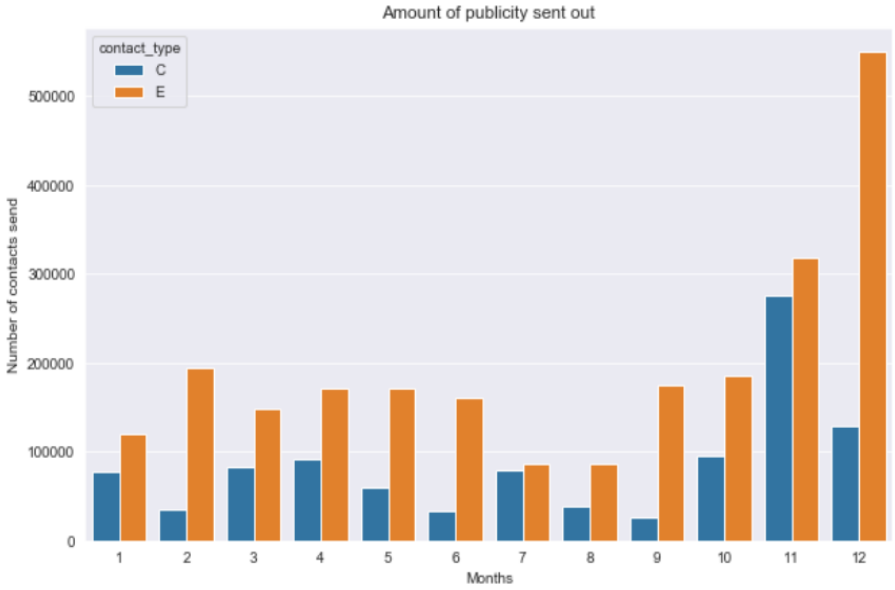

Figure 3: Sum of catalog and emails sent out per month 2005-2007

We can easily notice there is a high seasonality related to the core of the business around Christmas time. We would like to know on which month is our publicity the most effective. To define this parameter, we will consider the period of the same month where a customer was contacted and that the customer placed an order to investigate the periods where our customers are purchasing within the same time frame as the one that they received our marketing efforts.

```
SELECT *
FROM
    (SELECT cust_id,
        EXTRACT(YEAR FROM contact_date) AS year,
        EXTRACT(MONTH FROM contact_date) AS month,
        contact_type,
        COUNT(contact_type) AS contact_count
    FROM contacts
```

```
    GROUP BY cust_id,
        EXTRACT(YEAR FROM contact_date),
        EXTRACT(MONTH FROM contact_date), contact_type) AS t1
LEFT JOIN (
    SELECT cust_id,
        EXTRACT(YEAR FROM order_date) AS year,
        EXTRACT(MONTH FROM order_date) AS month,
        COUNT(order_num) AS orders_count
    FROM orders
    GROUP BY cust_id,
        EXTRACT(YEAR FROM order_date),
        EXTRACT(MONTH FROM order_date)) AS t2
USING(cust_id, year, month)
WHERE orders_count IS NOT NULL
```

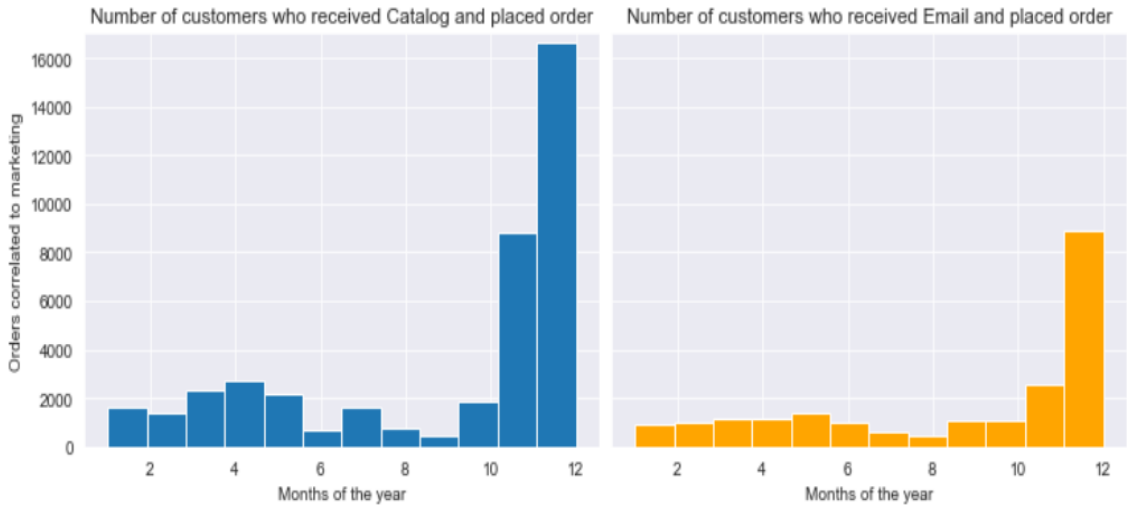

Figure 4: Comparison of customers that placed order after receiving catalogs vs emails

From this query we can extract the following graph which indicate, that customers that are receiving catalogs in the month of December are purchasing substantially more than the ones receiving emails for the last two months of the year. Purchases through emails is lower for November and December which can be attributed to the amount of publicity customers receive during the holiday period. The two major insights of this graph would be two consider saving money by not sending publicity in the months of June, July, and August which are extremely low in responses; Plus to send more catalogs during the months of March, April, May which appear to be effective at bolstering our sales during an off season period.

2.a Data-analytical Question 2: How does the percentage of gifts change over time?

We would like to investigate how the percentage of gifts has evolved over time. The company's core business consists of food products which are often purchased as gifts, it is crucial to understand how and when do this patterns of gift purchasing change during the year. We look at the overall picture of the different measures in the summary table of our customers purchases for the last 4 years to get an initial idea

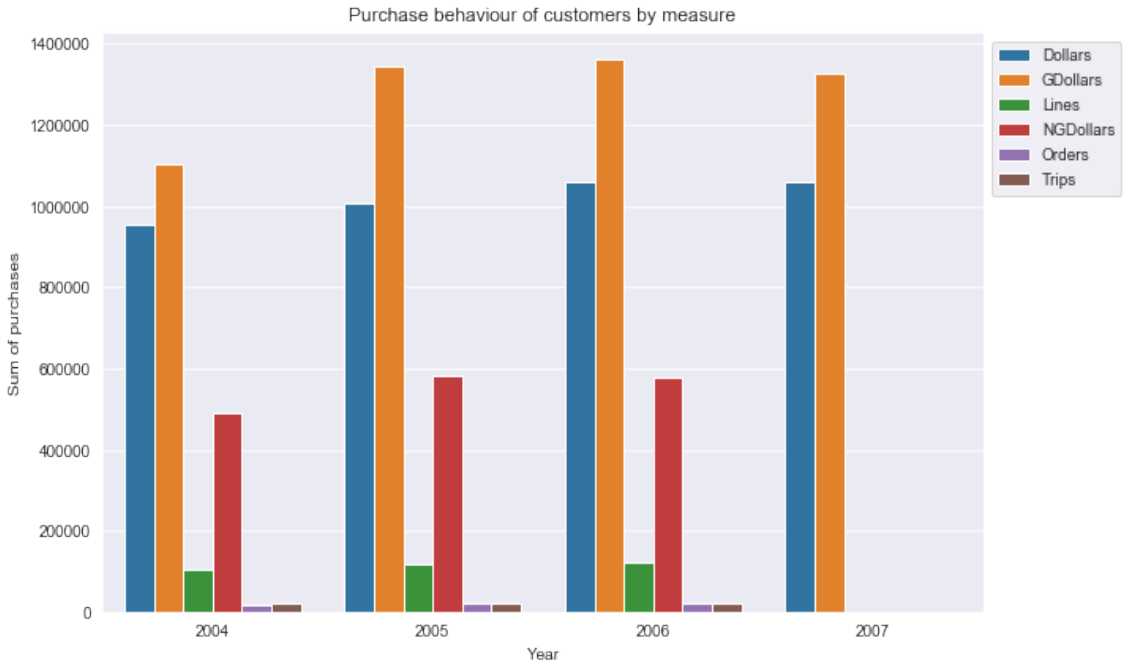

Figure 5: Comparison of the different purchases by method

Gift purchases are an important aspect of the sales, it is essential to understand from our customers that purchase our products as gift how do their patterns change over the months of the year. We can investigate this further by calculation the amount that our customers spent as a gift, divided by the amount that our customers purchased in total. We proceed by extracting the percentage of a customers order that is a gift per year and month.

```
SELECT cust_id,
    year,
    month,
    total_amount::numeric,
    gift_amount::numeric,
    contact_count,
    ROUND(gift_amount/total_amount*100) AS perc_gift
```

```
FROM (
SELECT cust_id,
    EXTRACT(YEAR FROM order_date) AS year,
    EXTRACT(MONTH FROM order_date) AS month,
    SUM(amount) AS total_amount
FROM lines
GROUP BY cust_id,
    EXTRACT(YEAR FROM order_date),
    EXTRACT(MONTH FROM order_date)) AS t1
LEFT JOIN (
    SELECT cust_id,
    EXTRACT(YEAR FROM order_date) AS year,
    EXTRACT(MONTH FROM order_date) AS month,
    SUM(amount) AS gift_amount
    FROM lines
    WHERE gift = 'Y'
    GROUP BY cust_id,
        EXTRACT(YEAR FROM order_date),
        EXTRACT(MONTH FROM order_date)) AS t2
USING (cust_id, year, month)
```

This query will give us the customers that have placed orders and we know that the order was categorized as gift on the lines table, which allows us to understand how much of their purchased is for themselves and how much is for other people. We create a new table as "Gifts" using the previous code and use the following query to extract the change of gifts percentage over time:

```
SELECT CONCAT(year, '-',month) AS date,
    year,
    ROUND(SUM(gift_amount)/SUM(total_amount),2 ) AS gifts_perc
FROM gifts
WHERE year <2008 AND year >=2004
GROUP BY year, month
ORDER BY year, month
```

We can now plot the data of our customer behavior for purchasing gifts for the last four seasons of available data, the information with changes of year over year is shown in the following graph:

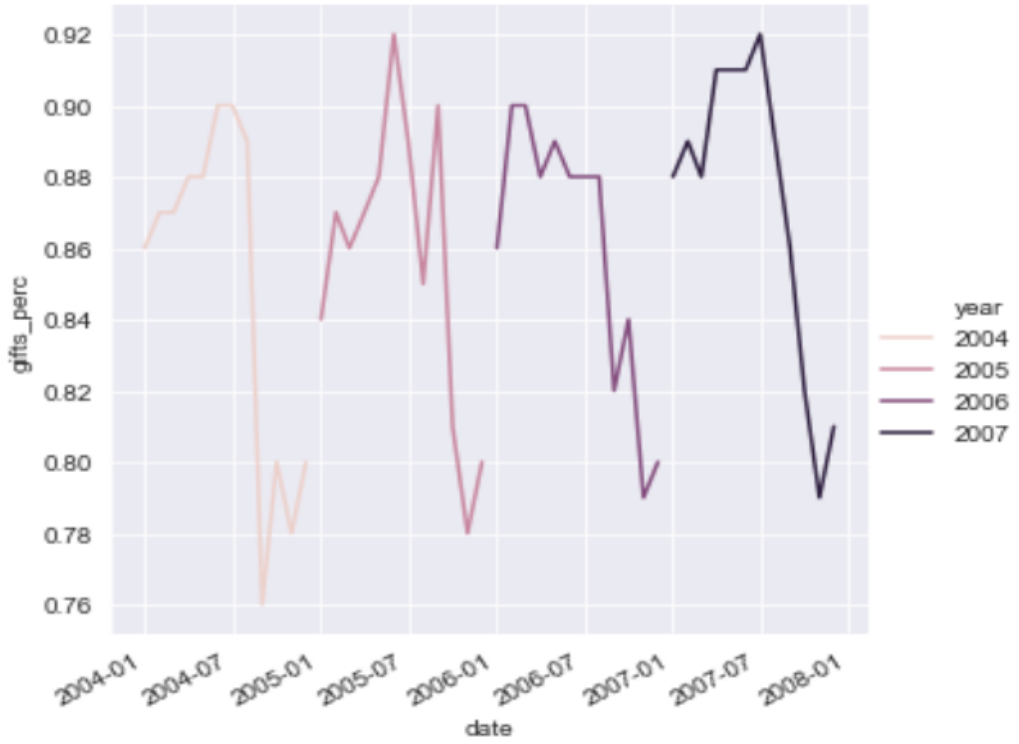

Figure 6: Evolution of customers purchase as percentage of gifts over time

There are some key insights in this graph, every year we can see a bigger progressing towards higher values of percentage of gifts purchased by customers. Although, this is the core business it is worth considering whether the company could plan a diversification effort to not become completely uni-dimensional. The second interesting aspect is the massive dip right after the Christmas period. There is a gap where the regular customer change the purpose of the purchase by roughly 10%, and the intention of the purchase is no longer a gift. It could be worth investigating the type of purchases that customers are making after the Christmas period which are not marked as gift to understand if there is a specific niche market for which the company's products are being bought.

2.b Segment using RFM dimensions

This section will show the segment of the customer data base using the RFM dimensions, it will be divided into 5 quantiles for each dimensions. Afterwards it will estimate the response rates for each RFM cell and show to how many segments the company should mail based on an estimate of ROI of the campaign.

```
CREATE TABLE rfm AS(
    SELECT cust_id,
         NTILE(5) OVER (ORDER BY last_order_date) AS recency,
         NTILE(5) OVER (ORDER BY count_order) AS frequency,
         NTILE(5) OVER (ORDER BY avg_amount) AS monetary
    FROM (
        SELECT cust_id,
          MAX(order_date) AS last_order_date,
          COUNT(*) AS count_order,
          AVG(amount::numeric) AS avg_amount
    FROM lines
    GROUP BY cust_id) AS table_1)
```

It is important to note that for the responses of the customers we are only considering the values from 2005 onward because our contacts data specifies that the marketing contact records only has values from 2005-2007.

```
CREATE TABLE rfm_rr AS(
SELECT rfm_segment,
    SUM(contacted) AS total_contacts,
    SUM(response) AS total_response,
    ROUND(SUM(response)/SUM(contacted),2) AS response_rate
FROM (
    SELECT cust_id, COUNT(*) AS contacted
    FROM contacts
    GROUP BY cust_id) AS t1
LEFT JOIN (
    SELECT cust_id, COUNT(*) AS response
    FROM orders
    WHERE order_date >= '2005-01-01'
    GROUP BY cust_id) AS t2
USING(cust_id)
LEFT JOIN (
    SELECT cust_id, CONCAT(recency, frequency, monetary) AS rfm_segment
    FROM rfm) AS t3
USING(cust_id)
```

```
GROUP BY rfm_segment)
```

We can now demonstrate the RFM segmentation and order by the response rate. This query allows us to easily see which segments have the best response rates of our customer pool. A normalized $1 per mailing and an average profit of $30 per purchase were kept as provided in the guidelines.

```
SELECT *,
    total_response*30 AS profit,
    total_contacts*1 AS cost,
    ROUND((total_response*30 - total_contacts*1)/total_contacts*1, 2) AS roi
FROM rfm_rr
WHERE response_rate IS NOT NULL
ORDER BY response_rate DESC
```

Data Output | Explain | Messages | Notifications

	rfm_segment text	total_contacts numeric	total_response numeric	response_rate numeric	profit numeric	cost numeric	roi numeric
1	551	17361	5952	0.34	178560	17361	9.29
2	451	20268	5736	0.28	172080	20268	7.49
3	431	2272	468	0.21	14040	2272	5.18
4	521	478	98	0.21	2940	478	5.15
5	341	6681	1210	0.18	36300	6681	4.43
6	351	15549	2765	0.18	82950	15549	4.33
7	441	5644	950	0.17	28500	5644	4.05
8	541	3135	528	0.17	15840	3135	4.05
9	552	54392	8649	0.16	259470	54392	3.77
10	452	53798	7924	0.15	237720	53798	3.42
11	331	4554	703	0.15	21090	4554	3.63
12	321	2723	346	0.13	10380	2723	2.81
13	531	1782	232	0.13	6960	1782	2.91
14	513	4968	578	0.12	17340	4968	2.49
15	512	864	101	0.12	3030	864	2.51
16	423	129	14	0.11	420	129	2.26

Figure 7: RFM segmentation with response rate and ROI

After analysing this data we notice that an appropriate threshold is to contact the customers segments that have a response rate above 3% because that is the threshold at which it was seen to have positive ROI. With this observation, we can filter our marketing efforts to only contact 50 RFM segments from the total customer population.

3. More Calculations

We continue our exploration of the data set to obtain further insights of our customer base.

3.a Calculate the Average CLV of a customer

For this section we will extract the average customer lifetime value. The year 2004 will be taken as a base year, and the following three year will be compared consecutively. For this required analysis it was specified that only the e-commerce setting was desired, therefore the analysis only includes the order methods where the category 'I' ='Internet' was present. For this calculation, the discount rate used was 10%.

```
SELECT *
FROM (
    SELECT EXTRACT(YEAR FROM order_date) AS year,
    COUNT(order_num) AS total_orders
    FROM orders
    WHERE order_method = 'I'
    GROUP BY year) as t1
LEFT JOIN (
    SELECT EXTRACT(YEAR FROM lines.order_date) AS year,
    ROUND(AVG(amount::numeric), 2) AS avg_amount
    FROM lines
    LEFT JOIN orders
    USING(order_num)
    WHERE order_method = 'I'
    GROUP BY year) as t2
USING(year)
```

Rate	10%			
Year	Total Orders	Avg Revenue Amount	Retention Rate	Adjusted Revenue
2004	7,786	£46.34	-	-
2005	10,706	£44.39	100%	£44.39
2006	10,434	£48.28	97%	£42.78
2007	10,445	£51.25	100%	£42.36
			CLV :	£129.52

Figure 8: Customer Lifetime Value for 'Internet' order method (period 2005-2007)

We conclude that for the period analyzed the average CLV was 129.52 pounds.

3.b Plan for calculating different CLV's

Finally, management believes different types of customers have different CLV's, we will delineate a a project that could obtain the necessary results that could be useful for the business to make future precise marketing efforts.

To obtain this model we would calculate the retention of our customers. We define retention if a customer made any purchase in the year, he/she is still a customer. As we have seen in the previous section, across all the customers over last three year period data, the average dollar amount of the initial order was approximately 42. We can set this as a threshold to make a hypothesis.

The hypothesis for this experiment would be, there exist customer whose customer initial purchase amount is above 42 and it could be predictive of overall bigger customer lifetime value. Likewise, the customers whose initial amount purchase was below 42, could therefore be regarded as having a lower customer lifetime value overall. Therefore, we are separating our customer base into 2 groups.

The difference in the average lifetime value between the two groups would be a great tool for the marketing department. It would allow them to make future marketing plans that increase the long term performance of the company. The customers that are identified on the basis of their lifetime value will bring in more profits in the future years and would be more responsive to marketing targeting. The potential for improved profits can be enacted by management in the following manner.

We calculate CLV of customers whose initial orders are partitioned by the 42 threshold. We can set up our table by aggregating the amount of repeated orders for the last couple of years and compare against the amount of the initial order. For this we will require to keep track of the two groups of customers since the base year and follow the churn rate to identify which segment of the two groups is coming back for repeated business.

If after comparing the the two groups we observed have a significant difference between both CLV for the time given, then we can conclude that that our customers long term value for the company can be predicted based on their initial purchase. This framework will allow the marketing department to allocate resources efficiently across customers and channels of communication.

Conclusion

This document has shown how to obtain valuable insights from company's data base that includes customer contacts, orders, and purchases. We covered how to load a database into PgAdmin, do exploratory data analysis, calculations for RFM and CLV, and finally outline a plan for finding different types of customer CLV's.