Analysis of US Ecommerce Record of 2020:

Office Supplies Category

1. **Executive summary**
2. **Introduction**
    - 2.1. About the Dataset
    - 2.2.Target audience and business problem
    - 2.3. Scope of the analysis
    - 2.4. Analysis methods and tools
    - 2.5. Preparation for the analysis
3. **Analysis**
        - 3.1. RFM for customer segmentation
        - 3.2. Pareto chart
        - 3.3. TPA (Traditional Profitability Assessment): Discounts management.
        - 3.4. Correlation
        - 3.5. Hypothesis testing
        - 3.6. Order buckets
4. **Summary and recommendations**
5. **Appendixes**

1. Executive summary

This report analyses product revenue and profitability in the **Office Supplies category** (company has 3 product categories: Office Supplies, Technologies and Furniture). The real company name is not disclosed, the dataset "**US E-commerce records of 2020**" is taken from **Kaggle** database. The Office Supplies category is the **biggest category** in terms of **distinct products** and the **most items sold**, and the **second biggest in terms of profit**. One year period data is analysed with a focus to the recent quarter (**Q4**). Focus group of this analysis is **Category managers,** who are responsible for adding/removing items to assortiment and making decisions on discounts strategy.

Problems to be analysed:

- Dropped Profit Margin in Q4, what affected that?
- Which are the best selling items? In terms of Revenue and Profitability
- Are there any "red flags"?
- What kind of products could be added to assortiment?

In addition to this report, a **dashboard** and **presentation** is made. The report describes all applied analysis methods in detail, such as: RFM customers segmentation, Pareto chart method, TPA (Traditional Profitability Assessment), correlation, hypothesis testing and order buckets analysis. While **dashboard** is made with intention to be used as **a tool for tracking the performance, product comparison, making decisions about discounts strategies and ect**. Dashboard consists of 4 pages: 1st- common information for all categories, and three separate pages for each category. And the **presentation** is intended for **Office Supplies category manager,** aiming to point out the most important findings with a focus on the recent quarter (Q4).

As already mentioned, the Office Supplies category contains the **biggest amount of distinct products** and the most items purchased are **up to 20 USD**, the **typical basket value is up to 500 USD** and contains **on average 5 items per order.** No surprise that **business customers tend to buy more expensive items** compared to consumer segment buyers. There are some items which are generating significant amount of revenue, however are **unprofitable** (eg. Fellowes PB500 Binding machine -56% Profit Margin and Ibico EPK-21 Binding system -155% margin in Q$), company should definitely monitor such products.

Profit Margin in Q4 dropped due to 60% increased discounts (compared to Q3), a goal for upcoming quarter should be to focus on not only the positive Revenue growth, but also strive to maintain positive **Profit Margin growth**, this could be achieved by:

- **Actively monitoring products with negative profit margin**, decide on a strategy how long a company can keep unprofitable products; remove/replace products which are unprofitable for a long time;
- Use **Pareto Chart** when making decisions **which products to prioritise**;
- Offer more **expensive products** to **Corporate segment** customers;
- **Extend the assortment** with **cheaper products** where the price is up to 20 USD.

2. Introduction

2.1. About the Dataset

In this analysis I used US E-commerce records of 2020 from Kaggle (https://www.kaggle.com/datasets/ammaraahmad/us-ecommerce-record-2020). Before starting the analysis, I explored the dataset which has 19 columns and 13 of them are discrete columns. The dataset contains data of one year sales and has in total 3,3K rows.

Unfortunately no information was provided about the company; that would allow me to provide more precise insights. Company is an US e-commerce retail business, it has 3 product categories: Technologies, Office Supplies and Furniture. Company's customers are widely spread across almost the whole country (in 47 states out of 50).

2.2. Business problem and Target audience

Since the dataset has several product dimensions such as categories, subcategories and products, I decided to start exploring the data from looking at the company's overall performance and then breaking it into **category dimensions**. I decided to stick to the **revenue and profitability** approach when analysing a company's performance and came up with a goal **to perform a product revenue and profitability analysis on a product category level.** Also to provide recommendations about best and worst performing products in each category.

The **target group** (or potential users) of the dashboard **is Product Category Managers** who are responsible for making decisions about an assortment in product categories (adding new products or removing); also deciding on discount strategies. The intention is that the dashboard would be constantly updated and reviewed with a purpose to track the results of each quarter.

2.3. Scope of analysis

The analysis consists of these parts:

1. A **Dashboard:**
    - 1st page: Total sales and comparison of 3 product categories;
    - 2nd page: Technologies Category
    - 3rd: Office Supplies Category
    - 4th: Furniture;
2. **Presentation**: Product Revenue and Profitability in Office Supplies Category
   ** Here I decided to pick one category and try to provide precise insights for the manager*

Since the scope of products and product subcategories is broad, I decided to focus mostly on the **Office Supplies category.** This written report is also mostly focused on this category. In this analysis, one year's performance is covered.

2.4. Analysis methods and tools

I used the following analysis methods (each method is covered more in depth in the “Analysis” part): TPA (Traditional Profitability Assessment), Pareto chart method, Price buckets, RFM, Correlation, Hypothesis testing. Also calculated KPIs such as: Net Profit Margin, Net Profit per Item, COGS.

The main tools used in the analysis were following:

- BigQuery (for SQL queries);
- Data Studio (for dashboard);
- Keynote (for presentation);
- PSPP (for correlation calculation, hypothesis testing);
- Statistical Analysis Tool extension for Google Sheets (for hypothesis testing);

2.5. Preparation for the analysis

Checked for null values, duplicates, data formats. There were No missing values or duplicates, the dataset seems clean and ready to use. I imported csv. file to BigQuery, then connected to DataStudio.

3. Analysis

I started the analysis parts from calculating the main **KPIs** (ref. *1st Appendix),* since the main focus of the analysis is Revenue and Profitability, I used the following KPIs: Net Revenue, Net Profit, Net Profit Margin, Quantity of items sold.

3.1. RFM customer segmentation

To perform RFM analysis, I divided customers into four equal groups according to distribution of values for **recency**, **frequency** and **monetary** values. For this purpose I wrote SQL query (ref. *2nd Appendix* for SQL query).

- *Recency* — how recent the last transaction is. We would want to continually engage recent buyers and discover why less recent buyers have lapsed.
- *Frequency* — how many times the customer has bought from us.
- *Monetary* — how much each customer has paid for our products and services.

The RFM Segmentation was executed using these four steps:

1. Compute for recency, frequency, and monetary values per customer
2. Determine quartiles for each RFM metric
3. Assign scores for each RFM metric
4. Define the RFM segments using the scores in step 3

After calculating recency, frequency and monetary values, I assigned scores from 1 to 4 for each value. Where for F and M we give higher scores for higher quintiles (1 means the highest monetary and frequency value) and R is reversed- more recent customers get a score of 4.

The next step was to define RFF segment; since we have 4 groups of each R, F and M metrics, there are 64 potential permutations (4 x 4 x 4), a number is too much to manage so I decided to use the approach suggested in DMA guide, which suggest to combine the scores in the frequency and monetary aspect in terms of averaging. By doing this way I defined R vs. FM scores.

Each customer have an RFM segment assignment like this

Row	Customer_ID	recency	frequency	monetary	r_score	f_score	m_score	fm_score	rfm_segment
1	BV-11245	336	1	207.0	1	1	2	2	Hibernating
2	FH-14350	213	2	958.336	1	3	4	4	Hibernating
3	CS-12355	7	2	959.922	4	3	4	4	Recent Customers
4	DB-13555	114	1	9.096	2	1	1	1	At Risk
5	AB-10255	167	2	482.954	2	3	3	3	At Risk
6	MG-17680	93	2	257.496	3	3	2	3	Active Customers
7	HG-15025	202	1	63.92	1	1	1	1	Hibernating
8	RS-19765	37	3	803.882000...	4	4	4	4	Recent Customers

Since the dataset consists of only one year data, and average frequency is 2, I decided not to use 11 segments as it is often done, instead I picked 4 segments and divided all customers based on recency score. First I tried different combinations of assigning FM scores among the segments, but then I plotted them on a bubble chart, some segments were overlapping so it was difficult to interpret the results.

RFM customer segments

Customer Segment	Recency Score Range	Frequency and Monetary Combined Score Range
Recent Customers	4	1-4
Active Customers	3	1-4
At Risk	2	1-4
Hibernating	1	1-4

RFM for Office Supplies Category

Using bubble charts as a method of visualisation enables us to display three types of information at a time: Monetary (sizes of the bubbles), Frequency (y-axis), and Recency (x-axis).

We can see that in **Office Supplies** Category the biggest segment in terms of Revenue is "**Recent Customers**" (83K USD Revenue), the size of "**Active Customers**" and "**At Risk**" segments is very similar - 62K USD, while "**Hibernating**" serment adds about 37K USD to Revenue.

Sources:
https://towardsdatascience.com/a-simple-way-to-segment-customers-using-google-bigquery-and-data-studio-f31c8896cc52
http://www.silota.com/docs/recipes/sql-recency-frequency-monetary-rfm-customer-analysis.html

3.2 Pareto Chart

The Pareto principle is used in many fields: from business to sociology, it tells us that for various cases, 80% of the consequences come from 20% of the causes. In this case, I aimed to identify **20% of products which generate 80% of revenue.**

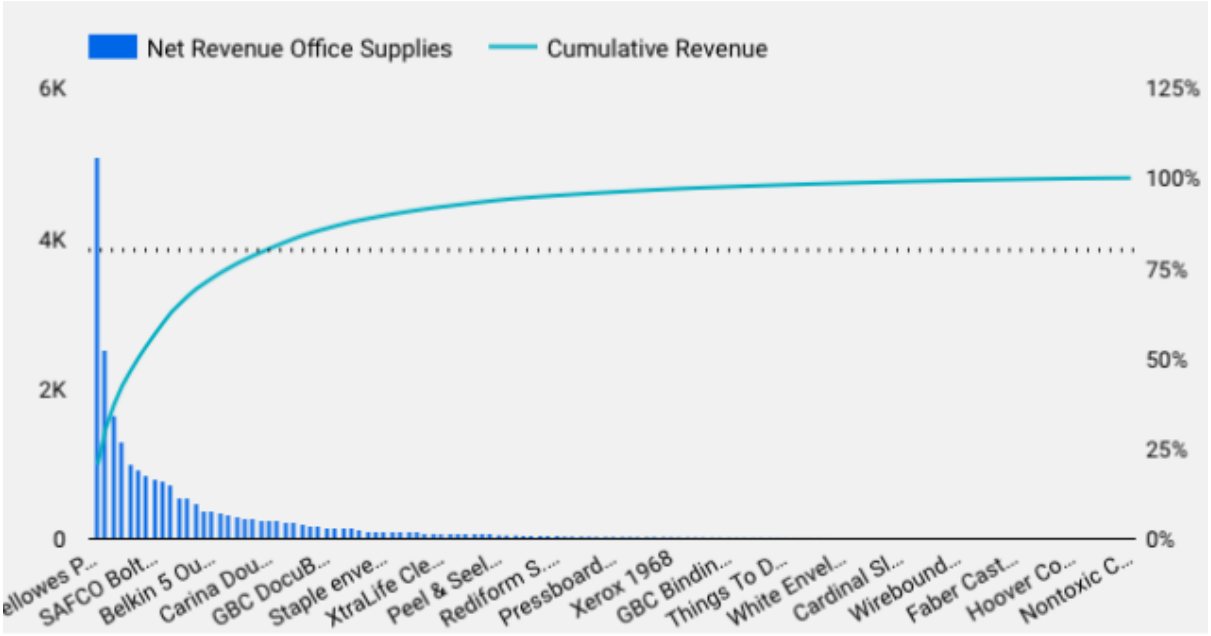
Pareto principle allows us to concentrate our efforts on those products that generate that 80% revenue. The main disadvantage of Pareto analysis is that it does not provide solutions to issues; it is only helpful for determining or identifying the root causes of a problem(s). In addition, Pareto analysis only focuses on past data.

In this case then I plotted **products and revenue**, it appeared that the company has a wide spread of products with **not many significant bars meaning that there is no clear Pareto pattern** and that there is no easy way to get breakthrough improvements.

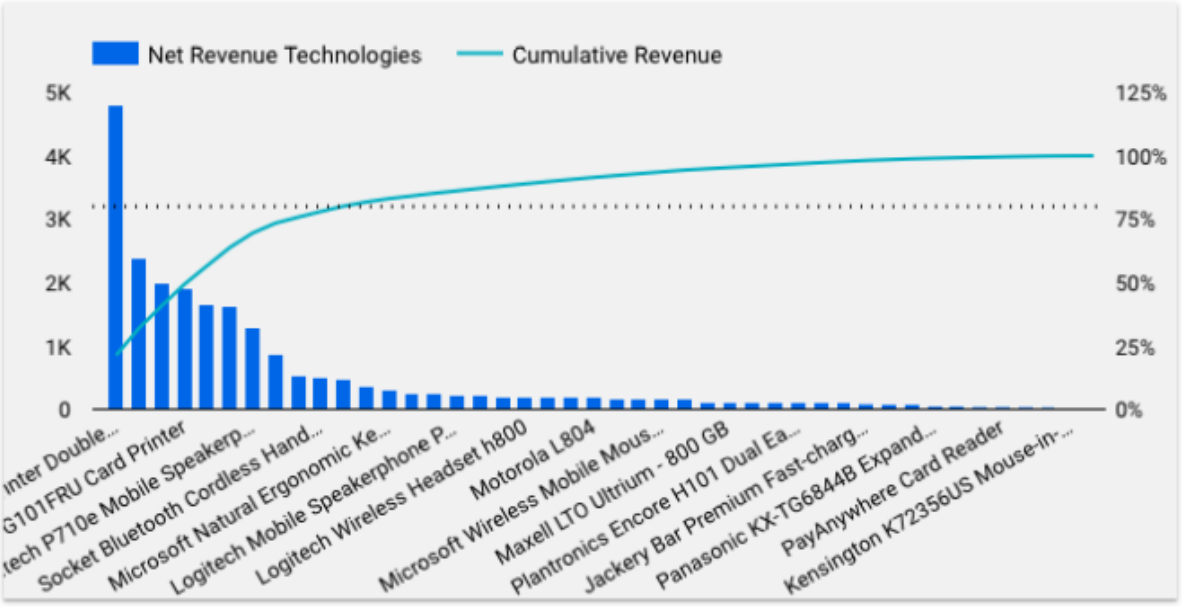
Pareto Chart **did not work neither for products, product subcategories nor for customers or specific customer segments.** There are more than 1500 distinct products and customers are mostly buying cheaper products. In most cases 80% of revenue was generated by 30% of products/ customers.

Pareto Chart method eventually **worked for the Office Supplies category in California state.** Also for the Technology category in California state (the biggest selling location) for the Corporate business segment.

Office Supplies category in California State. 68 out of 311 products generate 80% of Revenue:



Technologies category in California State in Corporate Segment. 10 out of 44 products generate 80% of Revenue



	Product_Name	Net_Reven...	Cumulative ...	Cumulative %
1.	Cubify CubeX 3D Printer Double Head Print	4.80K	4.80K	21.26%
2.	Hewlett Packard LaserJet 3310 Copier	2.40K	7.20K	31.89%
3.	Wilson Electronics DB Pro Signal Booster	2.00K	9.20K	40.77%
4.	Bady BDG101FRU Card Printer	1.92K	11.12K	49.27%
5.	Logitech Z-906 Speaker sys - home theater - 5...	1.65K	12.77K	56.58%
6.	Polycom SoundPoint IP 450 VoIP phone	1.63K	14.40K	63.79%
7.	Logitech P710e Mobile Speakerphone	1.29K	15.69K	69.49%
8.	Sharp 1540cs Digital Laser Copier	879.98	16.57K	73.39%
9.	StarTech.com 10/100 VDSL2 Ethernet Extend...	532.72	17.10K	75.74%
10.	Socket Bluetooth Cordless Hand Scanner (CH...	506.28	17.61K	77.99%

1 - 44 / 44

3.3. TPA (Traditional Profitability Assessment): Discounts management.

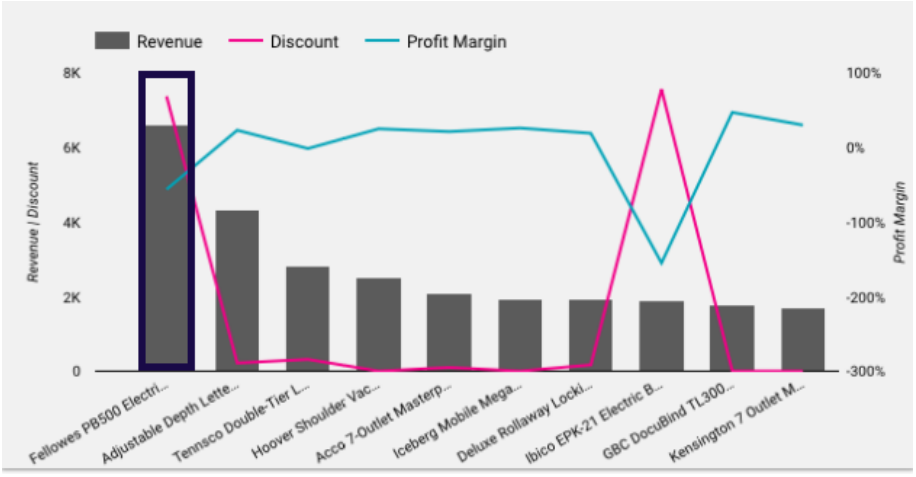
Traditional Profitability Assessment is very limited by the type of data available in the dataset. The dataset contains information about sales, revenue and discount rates, however there is no information provided about e.g. the shipping costs. From TPA only Discounts Management is partially covered in this analysis. Other potential assessments are listed below for further analysis.

Discount amount directly affects Revenue and Net Profit Margin. Discount amount KPI is added to best/ worst product charts in Data Studio dashboard in order to see whether increasing or decreasing of discount has a positive or negative impact on Revenue and Net Profit Margin. The primary intention of such charts in the dashboard is to monitor the products profitability taking into account the discount strategy and making changes if needed. Additionally, quarterly sales targets can be added.

Speaking about the company's **discounts strategy**, it is important to mention that **discounts are being made not for the whole order but for specific products.**

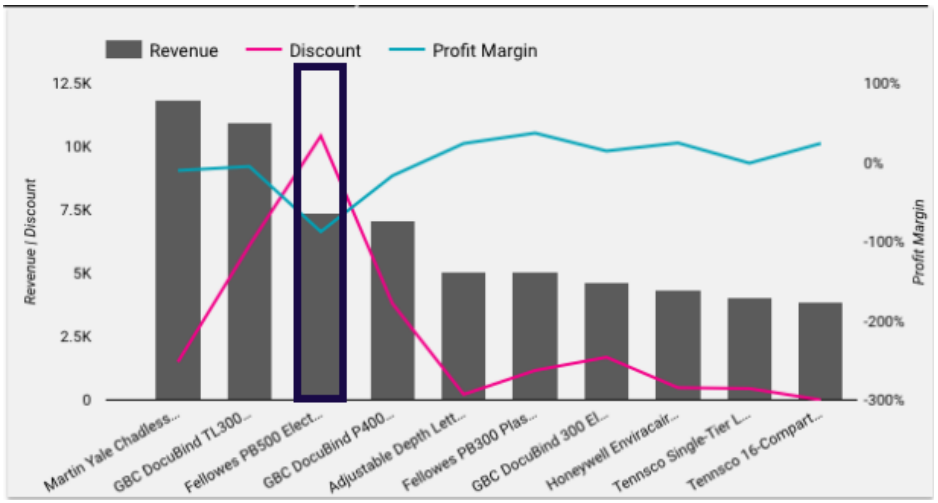
Let’s take an example, a product called Fellowes PB500 which in terms of Revenue was a Top performer in Q4, however a Profit Margin was -56% and company made discounts for 7.3K.

Q4 10 best products by Revenue



If we look at overall One-Year performance by Revenue, Fellowes PB500 is in TOP 3, looks promising, but if we look more closely at the Profit Margin, the result is even worse -87%!

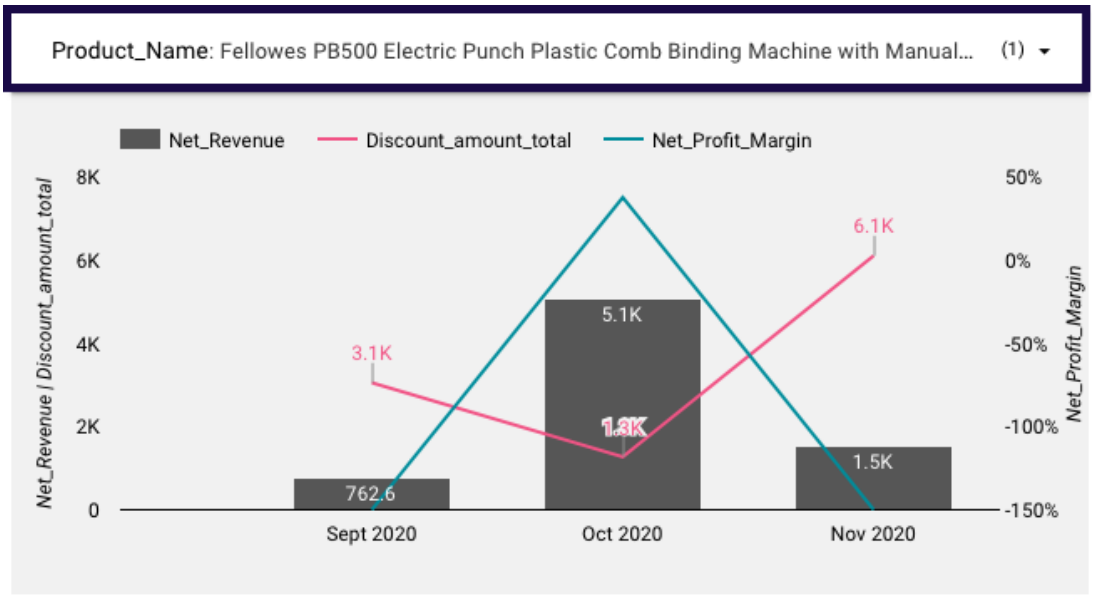
One Year 10 best products by Revenue



Ok, still not clear what is going on, let's look at the monthly graph. It looks like the product is quite new (first sales in September) and nothing was sold in December. Company sold 3 items in Sept, 5 in Oct and 6 in Nov. In October the Profit Margin was actually positive 37% and then dropped in November. It seems that the sales and discount strategy applied i**n October was successful** so it is worth considering repeating it if possible;

The strategy to increase discount **in November** increased the quantity of sold items (6 items, while in Oct 5 items were sold), but in terms of profit **it did not give positive results.** There are no sales in December, it is possible that the product was replaced with a different product, or nothing was sold. Company should think of ways to make this product more profitable or remove it from the assortment.

Monthly graph for Net revenue vs Discount amount vs Net Profit Margin



3.4. Correlation

Correlation Analysis is a statistical method that is used to discover if there is a relationship between two variables/datasets. For bivariate correlation I used the PSPP tool.

- Checking the correlation between **order value** and **amount of items**, it appeared that **weak correlation (0.259) exists** between those two variables.

Correlations

		Net_revenue_order	Quantity
Net_revenue_order	Pearson Correlation	1,000	,259
	Sig. (2-tailed)		,000
	N	526	526
Quantity	Pearson Correlation	,259	1,000
	Sig. (2-tailed)	,000	
	N	526	526

- Checking the correlation between **shopping frequency** (how many times customers made a purchase) and **order value**, we can say that there is **no correlation** (-0.022) between those two variables.

Correlations

		Net_revenue_order	Shopping_time
Net_revenue_order	Pearson Correlation	1,000	-,022
	Sig. (2-tailed)		,377
	N	1687	1687
Shopping_time	Pearson Correlation	-,022	1,000
	Sig. (2-tailed)	,377	
	N	1687	1687

- Also **no correlation** (-0.019) between **frequency** and **quantity of items** in the order.

Correlations

		Shopping_time	Quantity_order
Shopping_time	Pearson Correlation	1,000	-,019
	Sig. (2-tailed)		,447
	N	1687	1687
Quantity_order	Pearson Correlation	-,019	1,000
	Sig. (2-tailed)	,447	
	N	1687	1687

- No wonder that **correlation** exists **between revenue and quantity** of items bought (moderate 0.447 correlation); the more items in the basket, the more revenue the company earns.
  Also **correlation** exists **between revenue and discount amount** (not discount percentage- no correlation here), although correlation is weak (0.361). **Increasing the discount amount increases revenue.**

Finally, there is **no correlation** between **quantity of items and discount amount,** this can be explained by the fact that discounts are being made not for the whole order but for specific items.

Correlations

		Net_revenue_order	Discount_amount_total	Quantity_order
Net_revenue_order	Pearson Correlation	1,000	,361	,447
	Sig. (2-tailed)		,000	,000
	N	1722	1722	1722
Discount_amount_total	Pearson Correlation	,361	1,000	,052
	Sig. (2-tailed)	,000		,031
	N	1722	1722	1722
Quantity_order	Pearson Correlation	,447	,052	1,000
	Sig. (2-tailed)	,000	,031	
	N	1722	1722	1722

3.5. Hypothesis testing

In hypothesis testing part the intention was to check if behaviour of first time buyers is different from returning customers. For example if returning customers tend to make bigger orders or buy more items or maybe returning customers are buying more expensive products compared to first time buyers.

First Hypothesis

H_0: There is no difference between **first time buyers** and **returning** customers **order value**
H_1: There is a difference between **first time buyers** and **returning** customers **order value**

We cannot reject the H0 hypothesis, with 95% confidence level, when p value is > 0.05. Ref. to *3rd Appendix* for detailed Independent Samples T Test results.

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
order_value	Equal variances assumed	1,31	,253	1,02	1685,00	,306
	Equal variances not assumed			1,04	258,68	,297

Second Hypothesis

H_0: There is no difference between **first time buyers** and **returning** customers **size of order** (how many items in order)
H_1: There is a difference between **first time buyers** and **returning** customers **size of order**

We cannot reject the H0 hypothesis, with 95% confidence level, p value is > 0.05. Ref. to *4th Appendix* for detailed Independent Samples T Test results.

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
quantity	Equal variances assumed	4,35	,037	1,19	1685,00	,233
	Equal variances not assumed			1,12	246,17	,265

Third Hypothesis

H_0: There is no difference between **first time buyers** and **returning** customers **average item price**
H_1: There is a difference between **first time buyers** and **returning** customers **average item price**

We cannot reject the H0 hypothesis, with 95% confidence level, when p value is > 0.05. Ref. to *5th Appendix* for detailed Independent Samples T Test results.

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
Item_price	Equal variances assumed	,42	,519	,57	3310,00	,570
	Equal variances not assumed			,48	489,37	,631

It seems like there is no statistically significant difference in purchasing behaviour between these two groups: first time buyers and returning customers.

Fourth Hypothesis

H_0: There is no difference **Consumer** and **Corporate** segments **average item price**
H_1: There is a difference **Consumer** and **Corporate** segments **average item price**

With a confidence level of 95%, p value is < 0.05 (sig. 2-tailed 0.038), meaning that we can reject the H_0 hypothesis that there is no difference between those two means.
We can say that **Corporate clients buy more expensive products compared with the Consumer segment.** Mean in the Corporate segment is 83.09 USD, in the Consumer segment- 67.38 USD.

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
Item_price	Equal variances assumed	8,60	,003	-2,08	2646,00	,038
	Equal variances not assumed			-1,95	1671,74	,052

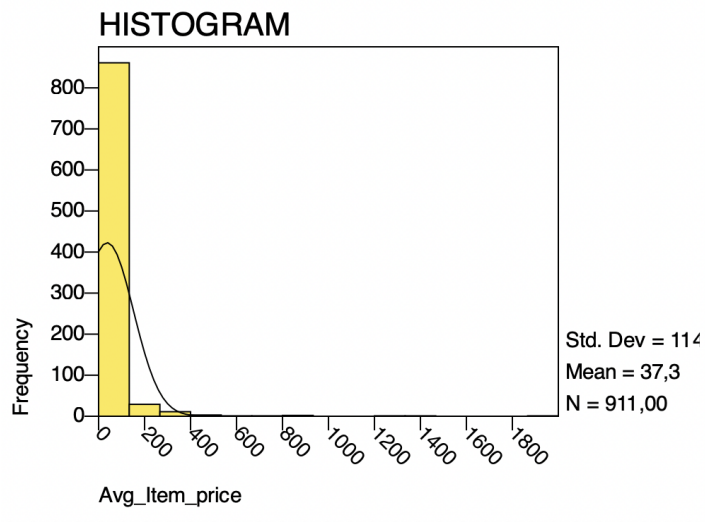
3.6. Order Buckets

Company has over 1.5K distinct products to offer, Office Supplies is the biggest category in terms of product variety (911 products in this category). I was interested to see what is the most common order bucket in the Office Supplies category.

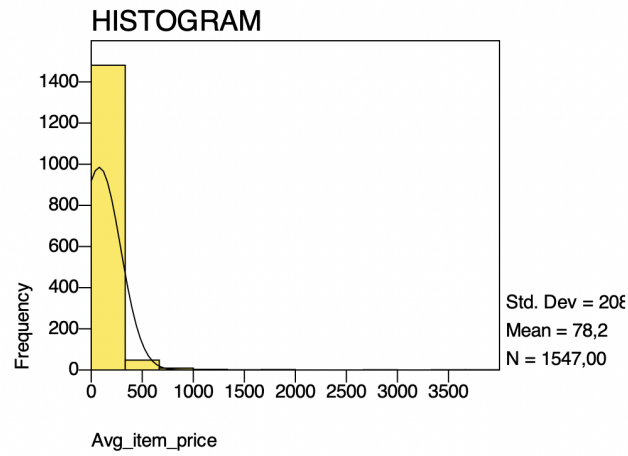
I started from checking minimum and maximum price ranges in category: minimum item price is 1.14 USD and maximum price is 1,890 USD, the average price is 37.3 USD.

Office Supplies category contains most distinctive products, which are cheaper - average price in all categories is 78.2 USD while in Office Supplies 37.3 USD (2 times lower)

Product price distribution in Office Supplies Category

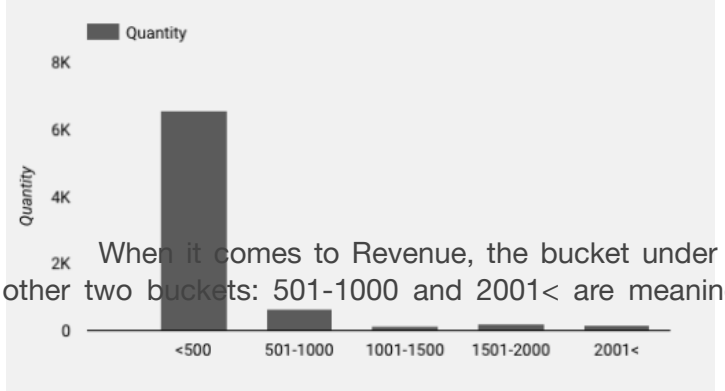


Product Price distribution in all categories



Next step was to divide orders into order buckets. I decided to use 500 USD intervals and used SQL to make these buckets (ref. to *7th Appendix* for SQL query). As we already know, most products in the Office Supplies category are under 100 USD, so no wonder that most products are sold when order value is under 500 USD.

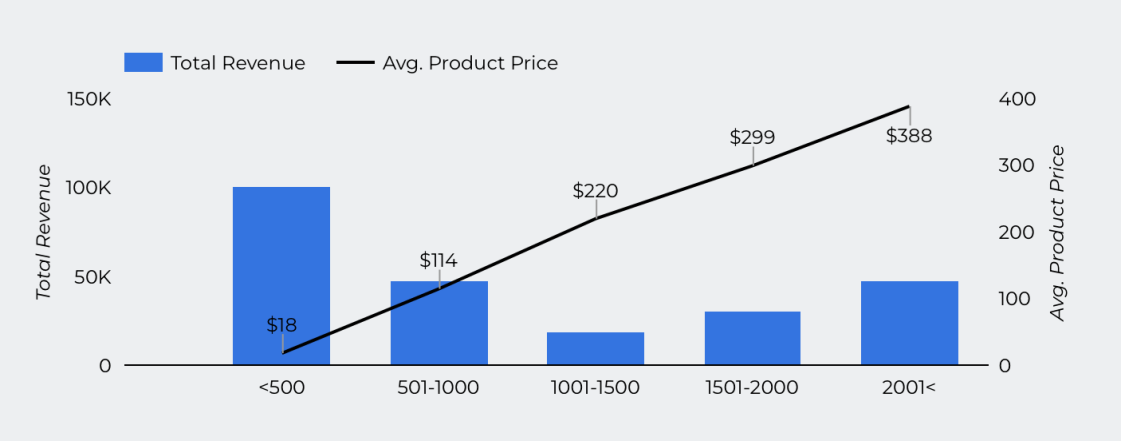
Quantity of items sold in Office Supplies category by Order Buckets



When it comes to Revenue, the bucket under 500 USD generates the most Revenue, but also other two buckets: 501-1000 and 2001< are meaningful. Average product price in the bucket which

generates the most revenue (<501 bucket) is 18 USD. It is obvious that customers buy more cheaper products from this category.

Revenue and Avg. Product price in Office Supplies category by Order Buckets



4. Summary and recommendations

- Office Supplies is the biggest category in terms of distinct products in the category and the quantity of items sold. **Category grew in Revenue by 18% in Q4** compared to Q3, however **Profit Margin dropped from 17.2% in Q3 to 13.4% in Q4**. One of the explanations for decreased Profit Margin is that the company made 60% more discounts compared to the previous quarter.

- RFM segmentation shows that the company does not have issues with attracting new customers- **"Recent Customers" is the biggest segment** in terms of Revenue, but it is worth paying more attention to the segment "At Risk" and put efforts not to lose these customers since this segment is as big as "Active Customers".

- Pareto Chart principle worked on the biggest market - California. It provides information about top products which generate the most revenue (20% of products generate 80% Revenue). Since Office Supplies category is very broad and it contains 911 distinct products, the intention of **Pareto Chart is to use it as a tool when making decisions which products to prioritise.**

- **TPA: discounts management analysis**, as mentioned previously, is quite limited due to insufficient data. Despite this limiting factor, I could see that there are some products (one example was mentioned previously- Fellowes PB500) where Revenue was growing, because the company applied huge discounts, however it resulted in a negative Profit Margin. The Company should pay attention to such "red flags".

Potential for further analysis:

Extend TPA (Traditional Profitability Assessment) adding:

- ➔ Geographical pricing (Are geographic price differentials justified? Is a customer eroding prices by ordering products from lower-priced geographic areas and having them shipped to higher-priced ones?)
- ➔ Scale pricing. Scale pricing should encourage customers to place fewer larger orders and cover any additional order (such as customer service representative and warehouse movements) or freight costs.
- ➔ Standard freight charges. Is the cost of freight a factor in setting prices? Does pricing reflect actual freight differentials? Look at pocket margin versus shipping zone for anomalies.
- ➔ Free, expedited freight. Are buyers charged for expedited freight when they request it?
- ➔ Variable payment terms. How many payment terms does the business use? Are customers paying late without penalty or receiving discounts beyond the term conditions? What is the business goal for accounts receivable?

- It appears that **correlation** exists **between revenue and discount amount** (not discount percentage- no correlation here), although correlation is weak (0.361). **Increasing the discount amount increases revenue.**

- Hypothesis testing between two customer segments: first time buyers and returning customers did not give expected results, there was no statistically significant difference among these two segments in terms of with order value, items per order or average item price was either very weak.

  However I found that **statistically significant difference is between Consumer and Corporate segments. Corporate clients buy more expensive products** compared with the Consumer segment customers.

- Order Buckets analysis showed that in the **most of orders are up to 500 USD** also such orders make up the biggest part of Revenue. Most products are sold when the average item price is 37.3 USD and **average order value is 87.4 USD.**

Recommendations

- **Actively monitoring products with negative profit margin**, decide on a strategy how long a company can keep unprofitable products; remove/replace products which are unprofitable for a long time;
- Use **Pareto Chart** when making decisions **which products to prioritise**;
- Offer more **expensive products** to **Corporate segment** customers;
- **Extend the assortment** with **cheaper products** where the price is up to 20 USD.

5. Appendices

1st. Appendix

Net Profit - from dataset
Net Revenue- from dataset (after discount, including shipping costs)
Discount Rate - from dataset
Net Profit Margin = Net Profit/ Net Revenue
Net Profit per Item = Net Profit/ Quantity
Discount Amount = ((Net Revenue/ (1-Discount Rate)/Quantity)- (Net Revenue/Quantity))
Average order value = Total Net Revenue / Total Count of Orders

2nd Appendix

SQL code:

Step 1: Compute for recency, frequency, and monetary values per customer

```
WITH t1
AS (
SELECT
DISTINCT(Customer_ID),
DATE_DIFF(Date('2020-12-31'), MAX(Order_Date), day) AS recency,
COUNT(DISTINCT(Order_ID)) as frequency,
SUM(Sales) AS monetary
FROM `training-363407.Capstone.US  E-commerce records 2020`
Group by 1
ORDER BY 1 desc),
t2 AS
```

Step 2: Determine quartiles for each RFM metric

```
(SELECT
  a.*,
  --All percentiles for MONETARY
  b.percentiles[offset(25)] AS m25,
  b.percentiles[offset(50)] AS m50,
  b.percentiles[offset(75)] AS m75,
  b.percentiles[offset(100)] AS m100,
  --All percentiles for FREQUENCY
  c.percentiles[offset(25)] AS f25,
  c.percentiles[offset(50)] AS f50,
  c.percentiles[offset(75)] AS f75,
  c.percentiles[offset(100)] AS f100,
  --All percentiles for RECENCY
  d.percentiles[offset(25)] AS r25,
  d.percentiles[offset(50)] AS r50,
  d.percentiles[offset(75)] AS r75,
  d.percentiles[offset(100)] AS r100
FROM
  t1 a,
  (SELECT APPROX_QUANTILES(monetary, 100) percentiles FROM
  t1) b,
  (SELECT APPROX_QUANTILES(frequency, 100) percentiles FROM
  t1) c,
  (SELECT APPROX_QUANTILES(recency, 100) percentiles FROM
  t1) d),
```

Step 3: Assign scores for each RFM metric

Now that we know how each customer fares relative to other customers in terms of RFM values, we can now assign scores from 1 to 4.

```
  t3 AS (
  SELECT *,
  CAST(ROUND((f_score + m_score) / 2, 0) AS INT64) AS fm_score
  FROM (
      SELECT *,
      CASE WHEN monetary <= m25 THEN 1
          WHEN monetary <= m50 AND monetary > m25 THEN 2
          WHEN monetary <= m75 AND monetary > m50 THEN 3
          WHEN monetary <= m100 AND monetary > m75 THEN 4
      END AS m_score,
      CASE WHEN frequency <= f25 THEN 1
          WHEN frequency <= f50 AND frequency > f25 THEN 2
          WHEN frequency <= f75 AND frequency > f50 THEN 3
          WHEN frequency <= f100 AND frequency > f75 THEN 4
      END AS f_score,
      --Recency scoring is reversed
      CASE WHEN recency <= r25 THEN 4
          WHEN recency <= r50 AND recency > r25 THEN 3
          WHEN recency <= r75 AND recency > r50 THEN 2
          WHEN recency <= r100 AND recency > r75 THEN 1
      END AS r_score,
      FROM t2
      )
),
```

Step 4: Define the RFM segments using the scores in step 3

```
t4 AS (
  SELECT
      Customer_ID,
      recency,
      frequency,
      monetary,
      r_score,
      f_score,
      m_score,
      fm_score,
     CASE WHEN(r_score = 3 AND fm_score = 2)
              OR (r_score = 3  AND fm_score = 1)  --Active Customers, At Risk, Recent
Customers, Hibernating
          OR (r_score = 3 AND fm_score = 3)
          OR (r_score = 3 AND fm_score = 4)
      THEN 'Active Customers'
      WHEN (r_score = 2 AND fm_score =3)
          OR (r_score = 2 AND fm_score = 4)
          OR (r_score = 2 AND fm_score = 1)
          OR (r_score = 2 AND fm_score = 2)
      THEN 'At Risk'
      WHEN (r_score = 4 AND fm_score = 1)
          OR (r_score = 4 AND fm_score = 4)
          OR (r_score = 4 AND fm_score = 3)
          OR (r_score = 4  AND fm_score = 2)
      THEN 'Recent Customers'
      WHEN r_score = 1 AND fm_score = 1
```

```
            OR (r_score = 1 AND fm_score = 2)
            OR (r_score = 1 AND fm_score = 3)
            OR (r_score = 1 AND fm_score = 4)
        THEN 'Hibernating'
        END AS rfm_segment
    FROM t3
) SELECT * from t4;
```

3rd Appendix

Group Statistics

	Group	N	Mean	Std. Deviation	S.E. Mean
order_value	First time buyer	200	495,58	875,98	61,94
	Returning	1487	426,43	898,99	23,31

Independent Samples Test

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
order_value	Equal variances assumed	1,31	,253	1,02	1685,00	,306	69,15	67,51	-63,25	201,55
	Equal variances not assumed			1,04	258,68	,297	69,15	66,18	-61,18	199,48

4th Appendix

Group Statistics

	Group	N	Mean	Std. Deviation	S.E. Mean
quantity	First time buyers	200	7,88	6,53	,46
	Returning	1487	7,33	5,99	,16

Independent Samples Test

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
quantity	Equal variances assumed	4,35	,037	1,19	1685,00	,233	,54	,46	-,35	1,44
	Equal variances not assumed			1,12	246,17	,265	,54	,49	-,42	1,50

5th Appendix

Group Statistics

	Group	N	Mean	Std. Deviation	S.E. Mean
Item_price	First time buyers	413	79,72	230,69	11,35
	Returning	2899	74,03	183,86	3,41

Independent Samples Test

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Item_price	Equal variances assumed	,42	,519	,57	3310,00	,570	5,69	10,01	-13,93	25,32
	Equal variances not assumed			,48	489,37	,631	5,69	11,85	-17,60	28,98

6th Appendix

Group Statistics

	Group	N	Mean	Std. Deviation	S.E. Mean
Item_price	Consumer	1668	67,38	168,45	4,12
	Corporate	980	83,09	217,02	6,93

Independent Samples Test

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Item_price	Equal variances assumed	8,60	,003	-2,08	2646,00	,038	-15,70	7,56	-30,53	-,87
	Equal variances not assumed			-1,95	1671,74	,052	-15,70	8,07	-31,52	,12

7th Appendix

```
SELECT
Order_id,
Order_Date,
Net_revenue_order,
Quantity,
IF(Net_revenue_order BETWEEN 0 AND 500, '<500',
IF(Net_revenue_order BETWEEN 500.01  AND 1000, '501-1000',
IF(Net_revenue_order BETWEEN 1000.01  AND 1500,'1001-1500',
IF(Net_revenue_order BETWEEN 1500.01 AND 2000,'1501-2000','2001<'))))AS Price_groups
FROM (
          SELECT
          DISTINCT(Order_id),
          Order_Date,
          SUM(Quantity) OVER (PARTITION BY Order_Id) AS Quantity,
          SUM(Sales) OVER (PARTITION BY Order_Id) AS Net_revenue_order
          FROM `training-363407.Capstone.US  E-commerce records 2020`
          WHERE Category = 'Office Supplies'
          ORDER BY 2 desc
          )
```