BY ELNARA & KATERINA

SUPERSTORE SALES

DATA PREPARETION PROJECT

About our dataset

We got a dataset about sales of a furniture § appliances store:
9994 rows x 21 columns
Features include: Order ID, Order Date, Ship Date, Customer Name, Segment, City, Sales, Quantity, Profit, etc

Project Goal

Apply the data preparation knowledge to clean the given dataset

Why do we need data cleaning?

Data cleansing, also known as data cleaning or scrubbing, identifies and fixes errors, duplicates, and irrelevant data from a raw dataset. Part of the data preparation process, data cleansing allows for accurate, defensible data that generates reliable visualizations, models, and business decisions.

Raw data

	A	B	C	D	E	F	G	H	I	J	K
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer II	Customer N	Segment	Country	City	State
2	1434	CA-2014-120768	12/19/2014	12/21/2014	Second Class	IM-15070	Irene Maddox	Consumer	United States	Florence	Alabama
3	1712	CA-2017-123491	10/30/2017	11/5/2017	Standard Class	JK-15205	Jamie Kunitz	Consumer	United States	San Francisco	California
4	7388	CA-2016-105732	9/13/2016	9/18/2016	Standard Class	AG-10270	Alejandro Grov	Consumer	United States	Omaha	Nebraska
5	8936	CA-2017-130036	8/27/2017	8/27/2017	Same Day	BP-11185	Ben Peterman	Corporate	United States	Philadelphia	Pennsylvania
6	2794	CA-2014-154599	4/12/2014	4/17/2014	Standard Class	KN-16450	Kean Nguyen	Corporate	United States	Redondo Beacl	California
7	6132	CA-2014-163447	12/27/2014	12/31/2014	Standard Class	TB-21190	Thomas Bruml	Home Office	United States	New York City	New York
8	661	CA-2015-146563	8/24/2015	8/28/2015	Standard Class	CB-12025	Cassandra Brar	Consumer	United States	Arlington	Texas
9	5386	CA-2017-161410	6/26/2017	7/3/2017	Standard Class	CC-12220	Chris Cortes	Consumer	United States	Philadelphia	Pennsylvania
10	9980	US-2016-103674	12/6/2016	12/10/2016	Standard Class	AP-10720	Anne Pryor	Home Office	United States	Los Angeles	California
11	9533	CA-2016-116596	10/27/2016	10/31/2016	Standard Class	BW-11200	Ben Wallace	Consumer	United States	New York City	New York
12	1430	US-2015-164448	10/31/2015	11/4/2015	Second Class	DK-12835	Damala Kotsor	Corporate	United States	Salinas	California
13	9169	CA-2016-140571	3/15/2016	3/19/2016	Standard Class	SJ-20125	Sanjit Jacobs	Home Office	United States	Jackson	Mississippi
14	6489	CA-2015-120621	3/21/2015	3/26/2015	Standard Class	JW-16075	Julia West	Consumer	United States	Jacksonville	North Carolin
15	6500	CA-2015-103135	7/24/2015	7/28/2015	Standard Class	SS-20515	Shirley Schmid	Home Office	United States	Louisville	Kentucky
16	6629	CA-2015-135489	9/19/2015	9/22/2015	Second Class	GW-14605	Giulietta Weim	Consumer	United States	New York City	New York
17	3903	US-2015-119312	10/2/2015	10/7/2015	Second Class	CS-12400	Christopher Sc	Home Office	United States	Los Angeles	California
18	1046	CA-2017-152702	10/12/2017	10/16/2017	Standard Class	SN-20710	Steve Nguyen	Home Office	United States	Rockford	Illinois
19	9942	CA-2017-164028	11/24/2017	11/30/2017	Standard Class	JL-15835	John Lee	Consumer	United States	San Francisco	California
20	5647	CA-2015-104241	1/4/2015	1/9/2015	Standard Class	AG-10495	Andrew Gjerts	Corporate	United States	Alexandria	Virginia
21	576	CA-2015-149713	9/18/2015	9/22/2015	Second Class	TG-21640	Trudy Glocke	Consumer	United States	Long Beach	California
22	3442	CA-2014-158337	3/11/2014	3/14/2014	Second Class	KA-16525	Kelly Andreada	Consumer	United States	New York City	New York
23	2948	CA-2017-169859	12/14/2017	12/18/2017	Standard Class	MP-18175	Mike Pelletier	Home Office	United States	San Diego	California
24	8522	CA-2016-141887	1/11/2016	1/15/2016	Standard Class	MP-17470	Mark Packer	Home Office	United States	Columbus	Ohio

Process

01 Import the data using Pandas

02 Examine the data for potential issues

03 Filling missing data

04 Removing duplicates

05 Handling with data inconsistency

06 Handling with outliers

07 Creating bins for profit measure

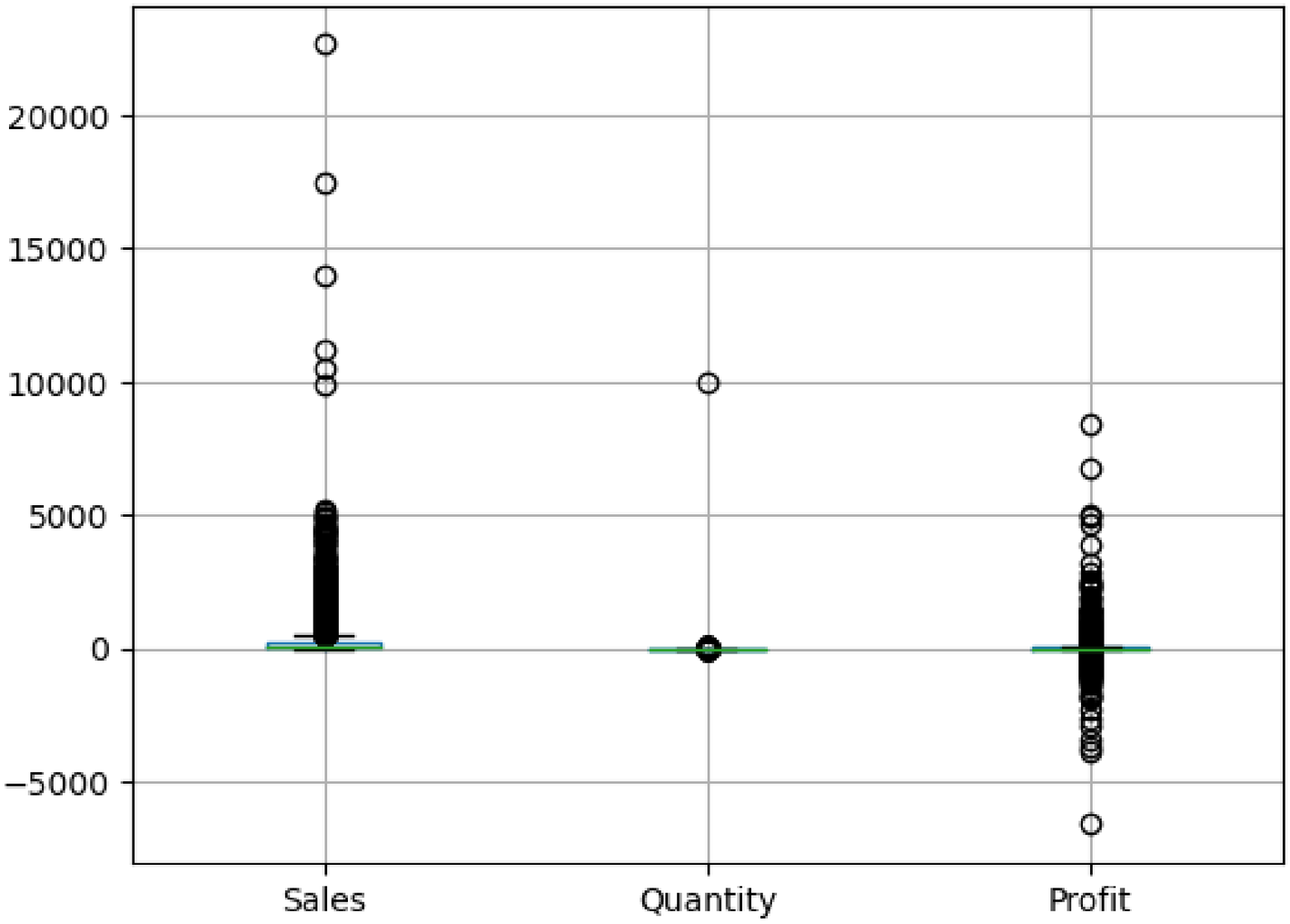
08 Identifying top 10 sales items and top 10 loyal clients

09 Visualisation of results

01 Exporting data into csv file

Highlights

After analyzing the outliers, no logical explanation or consistent pattern for these anomalies were detected, so we came to the conclusion that it was a type error that corrupts our data and we decided to remove 8 rows with a significant difference of almost two times from normal data.



Infographic About Sales

Outliers are clearly seen

Bins for profit measure

Customer ID	Customer Name	Segment	Country	City	...	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit	Profit
JK-15205	Jamie Kunitz	Consumer	United States	San Francisco	...	West	OFF-AP-10002684	Office Supplies	Appliances	Acco 7-Outlet Masterpiece Power Center, Wihtou...	1702.12	14	1.2	510.64	Very High
AG-10270	Alejandro Grove	Consumer	United States	Omaha	...	Central	FUR-FU-10003664	Furniture	Furnishings	Electrix Architect's Clamp-On Swing Arm Lamp, ...	1336.44	14	0.0	387.57	Very High
BP-11185	Ben Peterman	Corporate	United States	Philadelphia	...	East	TEC-AC-10001908	Technology	Accessories	Logitech Wireless Headset h800	1119.89	14	0.2	209.98	Very High
KN-16450	Kean Nguyen	Corporate	United States	Redondo Beach	...	West	TEC-PH-10001557	Technology	Phones	Pyle PMP37LED	1075.09	14	0.2	94.07	Very High
TB-21190	Thomas Brumley	Home Office	United States	New York City	...	East	FUR-CH-10004477	Furniture	Chairs	Global Push Button Manager's Chair, Indigo	767.21	14	0.1	161.97	Very High

Total of profit and sales per category

Sub-Category	Profit	Sales
Accessories	41936.73	167380.31
Appliances	18137.97	107532.14
Art	6527.77	27118.80
Binders	23630.70	167735.17
Bookcases	-3472.58	114879.97
Chairs	26426.31	326629.23
Copiers	36577.92	93028.24
Envelopes	6964.00	16476.38
Fasteners	949.52	3024.25
Furnishings	13059.19	91705.12
Labels	5546.18	12486.30
Machines	11795.78	121090.41
Paper	34052.92	78479.24
Phones	44516.03	330007.10
Storage	21278.83	223843.59
Supplies	-1189.12	38485.87
Tables	-17725.60	206965.52

Top 10 bestsellers

Sub-Category	Sales
Copiers	10499.97
Machines	5199.96
Binders	5083.96
Supplies	4912.59
Phones	4548.81
Chairs	4416.17
Bookcases	4404.90
Tables	4297.64
Accessories	3347.37
Storage	2934.33

Top 10 loyal clients

Customer Name	Sales
William Brown	37
Matt Abelman	34
John Lee	34
Paul Prost	34
Chloris Kastensmidt	32
Edward Hooks	32
Jonathan Doherty	32
Seth Vernon	32
Zuschuss Carroll	31
Arthur Prichep	31

MySQL Queries

```sql
select Category, round(sum(Sales)) as Total_Sales, round(sum(Profit)) as Total_Profit
from sales_project3.sales_clean
group by Category
order by Total_Profit desc;
```

Category	Total_Sales	Total_Profit
Technology	62867	12048
Office Supplies	49339	11345
Furniture	95666	5124

```sql
select Segment,  count(Sales) as All_Sales
From sales_project3.sales_clean
group by Segment
order by count(Sales) desc;
```

Segment	All_Sales
Consumer	166
Corporate	103
Home Office	63

```sql
select City, Region, round(count(Sales)) as Total_Sales
From sales_project3.sales_clean
group by Region, City
order by Total_Sales desc limit 10;
```

City	Region	Total_Sales
New York City	East	33
Los Angeles	West	25
San Francisco	West	18
Philadelphia	East	17
Seattle	West	17
Houston	Central	12
San Diego	West	11
Chicago	Central	11
Charlotte	South	5
Baltimore	East	5

Challenges

No consistent patterns to handle with outliers

Filling missing values

We spent much time searching for duplicates

Issues while creating SQL queries

Thank You For Attention! :)