Data Analysis

Bookstore Chain

Data Analyst Course - Project 4

Michael Orange

TABLE OF CONTENTS

01

Data Cleaning

Detect, correct and remove inaccurate and missing records

02

Data Analysis

Analyze data and identify patterns and trends for interpretation

03

Correlations

Investigate relationship between different variables of the business

01

Data Cleaning

Source datasets

CUSTOMERS

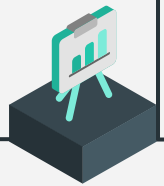
Client_id (key), sex, birth

len=8623

PRODUCTS

id_prod (key), price and category

len=3287



TRANSACTIONS

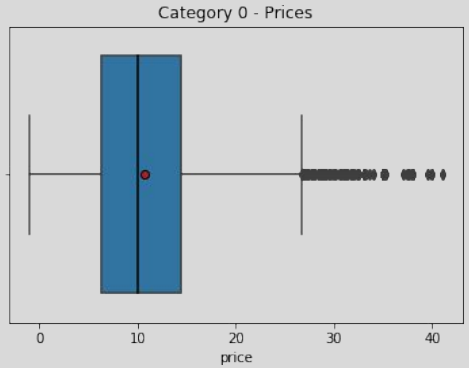
session_id, date, id_prod (key products), client_id (key cust)

len=337,016

Missing data

	id_prod	date	session_id	client_id	price	categ	sex	birth
18202	0_2245	2021-06-05 17:04:43.982913	s_44481	c_6714	NaN	NaN	f	1968.0
18203	0_2245	2021-04-22 04:57:20.090378	s_23987	c_6714	NaN	NaN	f	1968.0
20708	0_2245	2021-04-17 16:43:16.543156	s_21906	c_7808	NaN	NaN	m	1977.0
27016	0_2245	2021-11-01 14:00:26.535213	s_113219	c_7810	NaN	NaN	m	1978.0
30498	0_2245	2021-09-11 10:52:05.205583	s_88251	c_3468	NaN	NaN	f	1981.0
...	...	...	...	...	...	...	...	...
299794	0_2245	2021-09-19 03:08:45.918021	s_92049	c_4935	NaN	NaN	f	1982.0
311031	0_2245	2021-11-12 09:25:22.905886	s_118458	c_7416	NaN	NaN	m	1933.0
311482	0_2245	2021-06-17 03:03:12.668129	s_49705	c_1533	NaN	NaN	m	1972.0
324186	0_2245	2021-11-20 20:21:06.505658	s_122593	c_8524	NaN	NaN	f	1982.0
325475	0_2245	2021-05-20 07:44:21.415061	s_36985	c_1450	NaN	NaN	f	1959.0

103 rows × 8 columns



Product 0_2245 has no price and no category**.**

- id_prod code starts by '0_' which is the designation for the category 0.
- Price :
  - Mode is not a valid option : outside of the inter-quartile
  - Mean could be a valid option but due to the number of outliers, a **median substitution** is applied: **10.64**

Unwanted observations

	id_prod	date	session_id	client_id	price	categ	sex	birth
336768	T_0	test_2021-03-01 02:30:02.237420	s_0	ct_1	-1.0	0.0	m	2001.0
336769	T_0	test_2021-03-01 02:30:02.237446	s_0	ct_1	-1.0	0.0	m	2001.0
336770	T_0	test_2021-03-01 02:30:02.237414	s_0	ct_1	-1.0	0.0	m	2001.0
336771	T_0	test_2021-03-01 02:30:02.237434	s_0	ct_1	-1.0	0.0	m	2001.0
336772	T_0	test_2021-03-01 02:30:02.237412	s_0	ct_1	-1.0	0.0	m	2001.0
...	...	...	...	...	...	...	...	...
336963	T_0	test_2021-03-01 02:30:02.237437	s_0	ct_0	-1.0	0.0	f	2001.0
336964	T_0	test_2021-03-01 02:30:02.237438	s_0	ct_0	-1.0	0.0	f	2001.0
336965	T_0	test_2021-03-01 02:30:02.237436	s_0	ct_0	-1.0	0.0	f	2001.0
336966	T_0	test_2021-03-01 02:30:02.237445	s_0	ct_0	-1.0	0.0	f	2001.0
336967	T_0	test_2021-03-01 02:30:02.237430	s_0	ct_0	-1.0	0.0	f	2001.0

200 rows × 8 columns

200 transactions with a price under 0 from customers ct_1 and ct_2

- client_id are 'ct_' instead of standardized names 'c_', an additional 't' (as 'test').
- there is a flag 'test_' in the date description for these transactions.
- products are 'T-0' instead of the '0_', '1_' or '2_'.
- sessions are flagged 's_0'

ct_1, ct_2, T_0 and s_0 are **test accounts** and removed from the dataset.

Standardization and preparation

Orders

	session_id	age_bins	birth	client_age	client_id	date	sex	order_month	order_value
0	s_1	(53, 59]	1967.0	55	c_329	2021-03-01 00:01:07.843138	f	3	11.99
1	s_10	(47, 53]	1970.0	52	c_2218	2021-03-01 00:10:33.163037	f	3	26.99
2	s_100	(41, 47]	1978.0	44	c_3854	2021-03-01 04:12:43.572994	f	3	33.72
3	s_1000	(29, 35]	1989.0	33	c_1014	2021-03-03 02:49:03.169115	m	3	39.22
4	s_10000	(29, 35]	1989.0	33	c_476	2021-03-22 18:15:03.831240	f	3	41.49
...	...	...	...	...	...	...	...	...	...
169189	s_99994	(35, 41]	1983.0	39	c_7685	2021-10-04 18:56:23.112236	m	10	28.92
169190	s_99995	(59, 65]	1960.0	62	c_4170	2021-10-04 18:35:32.201073	f	10	19.84
169191	s_99996	(47, 53]	1974.0	48	c_4900	2021-10-04 18:39:10.485474	f	10	56.27
169192	s_99997	(41, 47]	1979.0	43	c_3521	2021-10-04 18:45:38.003516	f	10	6.99
169193	s_99998	(41, 47]	1978.0	44	c_2795	2021-10-04 18:45:54.374885	f	10	35.11

169194 rows × 9 columns

Customers

	client_id	age_bins	client_age	sex	sales	nb_orders	nb_prod	date_first_order	month_first_order	active_days	active_month	avg_order_value
0	c_1	(65, 71]	67	m	300.65	15	20	2021-06-11 21:02:39.382765	2021-06	262 days 02:57:18.657707	8.733333	20.043333
1	c_10	(65, 71]	66	m	586.18	16	28	2021-03-21 02:50:16.551727	2021-03	344 days 21:09:41.488745	11.466667	36.636250
2	c_100	(29, 35]	30	m	222.87	3	6	2021-04-20 05:26:43.133053	2021-04	314 days 18:33:14.907419	10.466667	74.290000
3	c_1000	(53, 59]	56	f	980.02	42	56	2021-03-13 13:34:14.618755	2021-03	352 days 10:25:43.421717	11.733333	23.333810
4	c_1001	(35, 41]	40	m	1102.45	24	58	2021-03-07 13:01:15.964197	2021-03	358 days 10:58:42.076275	11.933333	45.935417
...	...	...	...	...	...	...	...	...	...	...	...	...
8616	c_995	(65, 71]	67	m	119.08	5	8	2021-04-08 02:50:33.499557	2021-04	326 days 21:09:24.540915	10.866667	23.816000
8617	c_996	(47, 53]	52	f	739.60	36	42	2021-03-01 15:06:18.594244	2021-03	364 days 08:53:35.446228	12.133333	20.544444
8618	c_997	(23, 29]	28	f	572.89	10	23	2021-04-30 19:41:32.340325	2021-04	304 days 04:18:25.700147	10.133333	57.289000
8619	c_998	(17, 23]	21	m	1527.69	13	28	2021-03-18 01:32:33.677785	2021-03	347 days 22:27:24.362687	11.566667	117.514615
8620	c_999	(53, 59]	58	m	305.00	21	22	2021-07-10 23:59:59.414026	2021-07	232 days 23:59:58.626446	7.733333	14.523810

8621 rows × 14 columns

Products

	id_prod	categ	price	nb_sold	sales
0	0_0	0	3.75	611	2291.25
1	0_1	0	10.99	249	2736.51
2	0_10	0	17.95	12	215.40
3	0_100	0	20.60	2	41.20
4	0_1000	0	6.84	222	1518.48
...	...	...	...	...	...
3282	2_95	2	98.99	3	296.97
3283	2_96	2	47.91	281	13462.71
3284	2_97	2	160.99	5	804.95
3285	2_98	2	149.74	1	149.74
3286	2_99	2	84.99	2	169.98

3287 rows × 5 columns

Transactions

	id_prod	date	session_id	client_id	price	categ	sex	birth	client_age	age_bins
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	4.99	0	f	1977.0	45	(41, 47]
1	0_1085	2021-09-29 11:14:59.793823	s_97382	c_4450	3.99	0	f	1977.0	45	(41, 47]
2	0_1453	2021-08-27 19:50:46.796939	s_81509	c_4450	7.99	0	f	1977.0	45	(41, 47]
3	0_1405	2021-08-27 20:07:25.878440	s_81509	c_4450	4.99	0	f	1977.0	45	(41, 47]
4	0_1392	2021-12-28 11:45:04.072281	s_141302	c_4450	6.30	0	f	1977.0	45	(41, 47]
...	...	...	...	...	...	...	...	...	...	...
337011	1_607	2021-09-25 07:26:00.224331	s_95185	c_4786	26.99	1	f	1967.0	55	(53, 59]
337012	1_673	2021-06-01 00:49:49.781631	s_42350	c_2793	12.99	1	m	1933.0	89	(83, 89]
337013	0_2075	2021-10-09 09:03:48.268536	s_102200	c_2793	8.99	0	m	1933.0	89	(83, 89]
337014	0_1692	2021-09-15 19:42:08.596375	s_90430	c_4478	13.36	0	f	1970.0	52	(47, 53]
337015	0_142	2021-09-25 18:07:25.880052	s_95415	c_1232	19.85	0	f	1960.0	62	(59, 65]

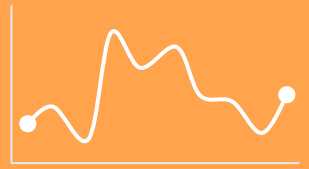
336816 rows × 10 columns

Exported as CSV to:

- data_orders.csv
- data_customers.csv
- data_prod.csv
- data_transac.csv

02

Data Analysis



- Customers
- Products
- Transactions

Bookstore Chain

Revenues
5,797,706

Orders
169,194

Products
3,287 references

Products Sold
336,816

Period
March 2021 to February 2022

Customers
8,621

Source dataset incomplete for the category 1

No orders from October 2nd to 27th for the category 1.

- Considering the number of references, shortage of stock is not a plausible hypothesis.
- It is certainly due to **missing data** in the source dataset.

It is important to **acknowledge that information for the rest of the analysis**.

An active and loyal base of customers

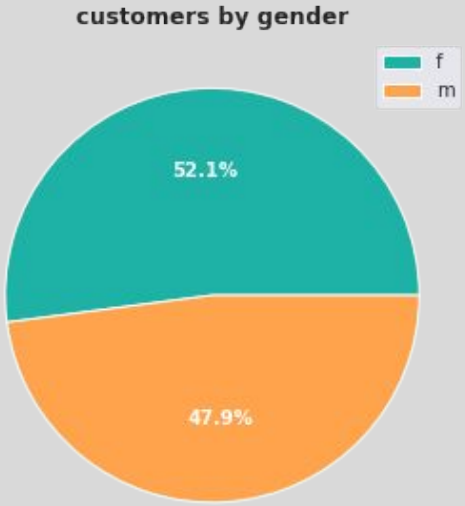
8,621 customers

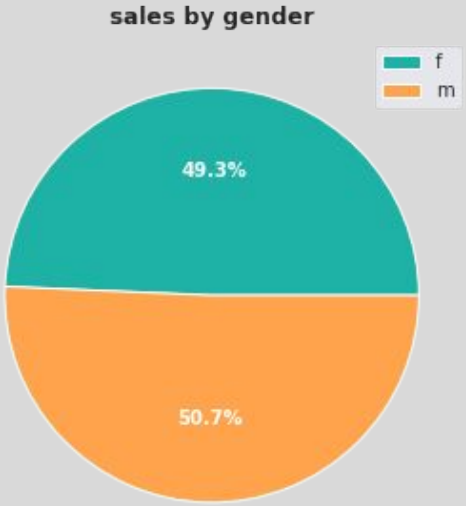
- 8,600 active customers
- 21 inactive customers

826.10 Customer value

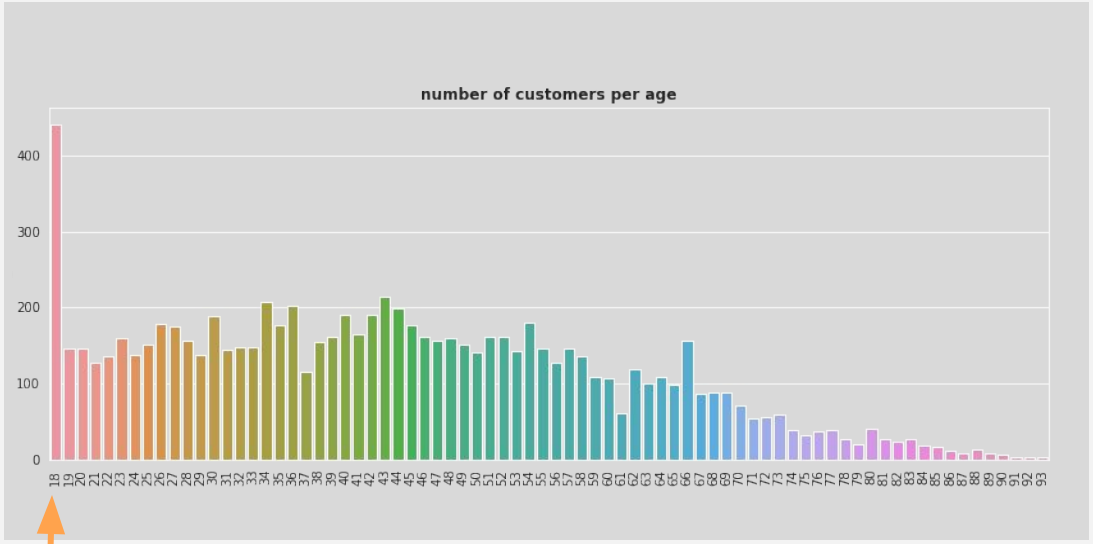
- 40.11 Avg Order Value
- 1.72 order/month
- -0.24% Churn rate

A balanced mix between genders





An Age-Diverse customer base



Age 18 is significant: it might be the default age in the system (ex. customers under 18 or customers who didn't fill their ages during the order process).

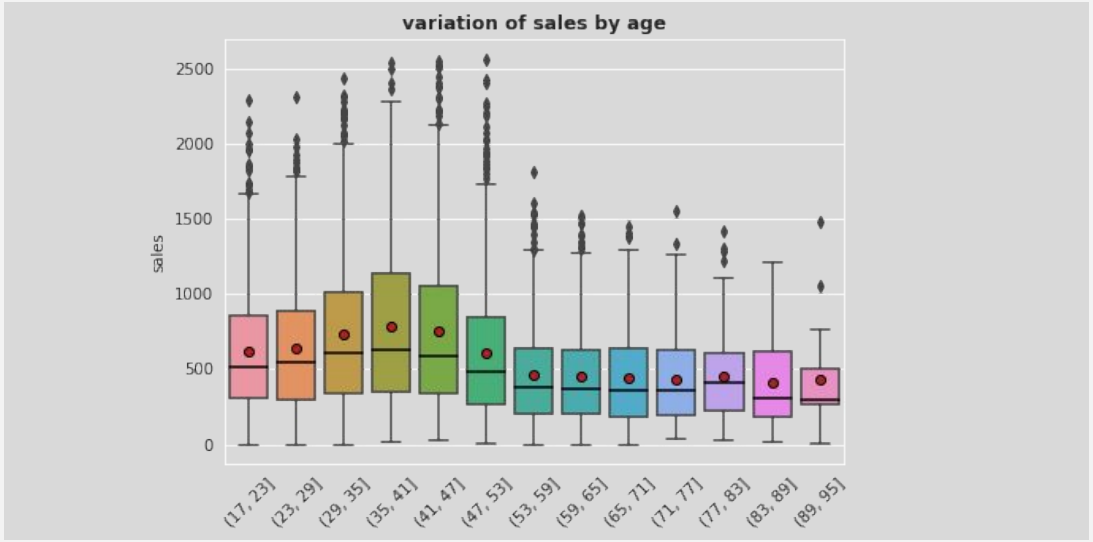
Customers from 18 to 93 years

Mean age: **43.75 years**

Median age: 43 years

Approx 70% are aged between 27 and 60 years

Customers aged of 30-47 are spending more



Customers aged between 35 and 41 are spending more compared to other categories.

Skewed right distribution of the sales per age categories.

Older customers are spending less.

50% of the sales from 20% of the customers

7,5%

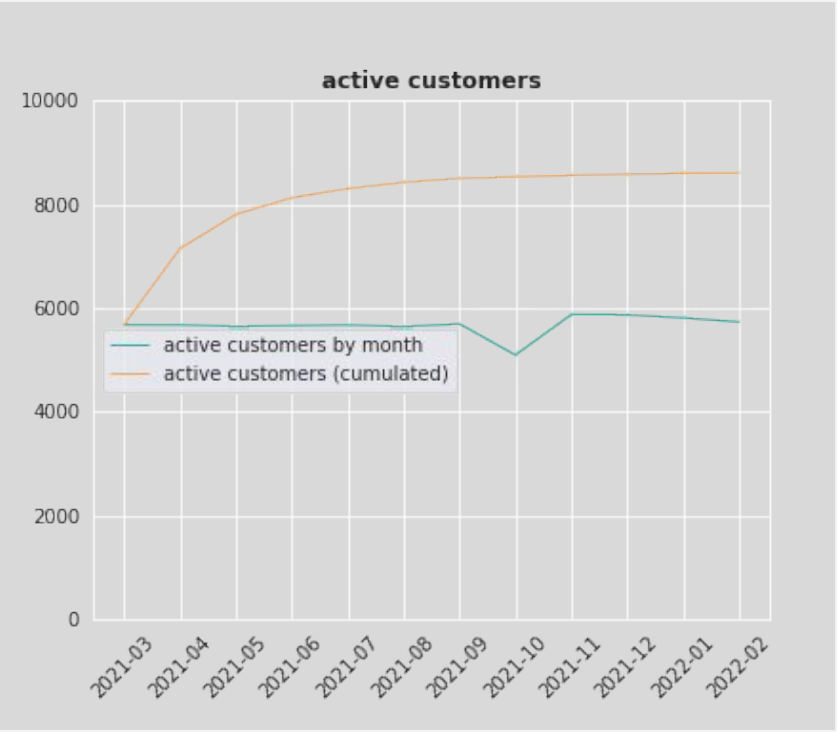
4 exceptional customers

Average of **108,486** of sales per year

Average of **2,846** orders per year

Might be communities (public libraries, schools, nursing homes, etc.)

Almost no recruitment of new customers



98 out of 100 customers in H2 are already existing customers

Only 178 net new customer in the last 6 months (H2)

99% **of customers buy more than once (repeat customers)**

A large catalog of products

336,8616 products sold

- 21.86 Avg Order Price
- 1,763 Average Revenue per product

3,287 products

- 3,265 active products
- 22 products not sold
- 2,309 in category 0
  739 in category 1
  239 in category 2

60%-80% of sales are done with 20% products

30% sales from 3% products

100 best-selling products

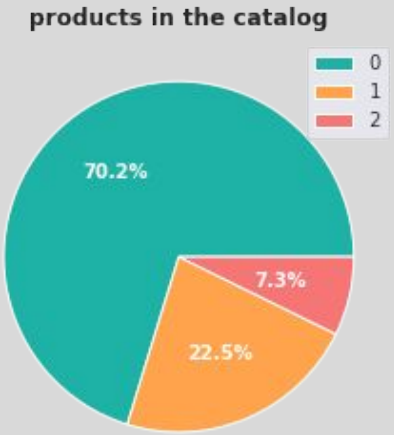
Revenue by product: **16,897**

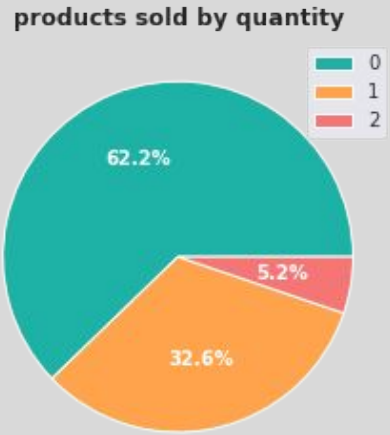
3% sales from 30% products

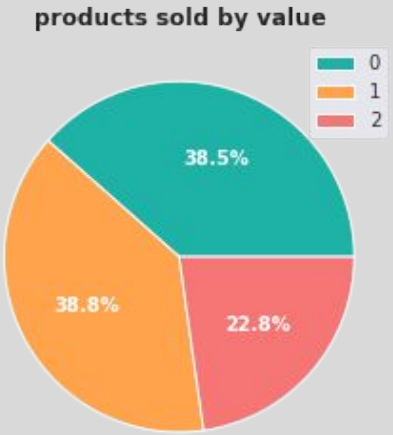
1,007 products (769 from categ 0)

Revenue by product: **129**

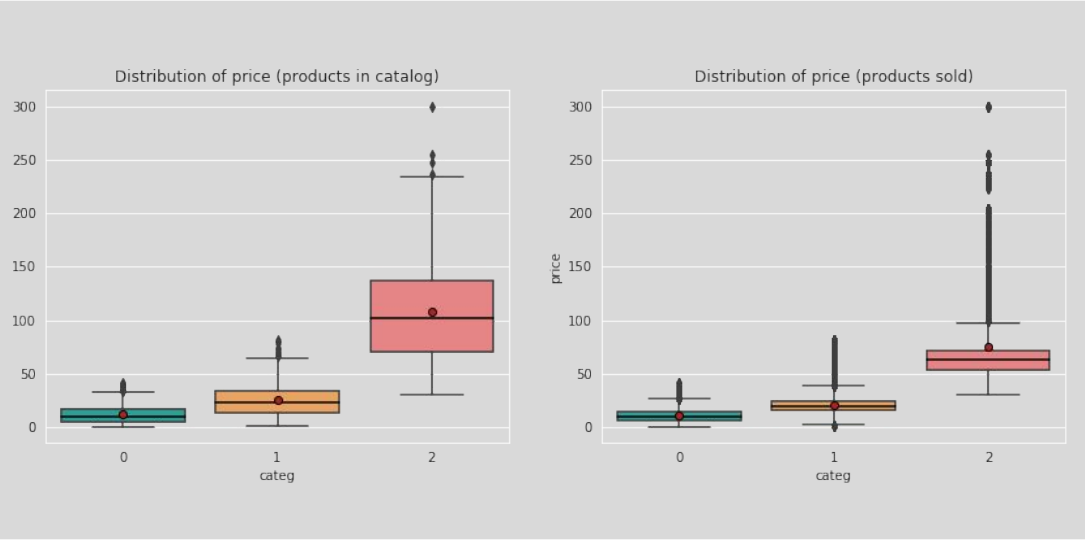
A balanced mix in terms of sales







Lower-price products in each category are more sold



- Category 0 avg price is **11.73**
- Category 2 avg price is **25.53**
- Category 3 avg price is **108.35**

Different price distribution if we consider the product actual sales.

Inner-quartile range and upper-quartile range are significantly reduced (with lower means).

The higher-price products of the catalog have lower volume of sales.

No growth to manage

Excluding the 2 months impacted by the lack of data in october:

0.98% of average growth per month

Categ 0: **1.7%**

Categ 1: **-0.96%**

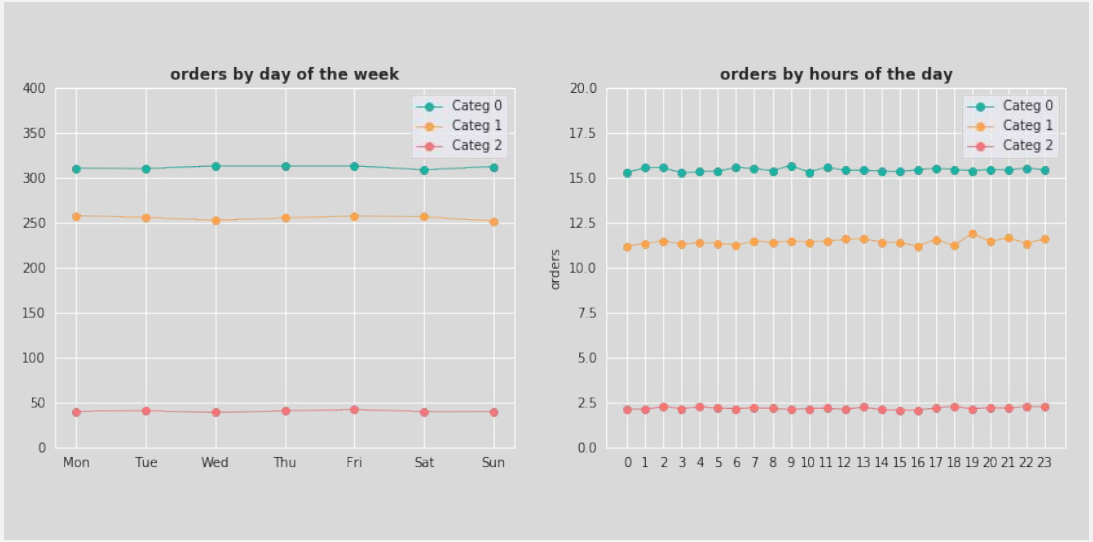
Categ 2: **8.37%**

A predictable business

Average order value remains stable with a mean of **34.16 per order.**

Average number of products per order remains stable with a mean of **1.99** products per order.

Uninterrupted orders 7 days a week and 24 hours a day



Number orders by day of the week remains constant.

Number of orders by hours of the day remains constant.

It needs to be investigated with bookstore chain management (ex. international sales)

03

Correlations

Outliers

Top 4 customers

They are not individuals and their individual characteristics are not representative.

They are excluded for the following statistical tests.

18 years old customers

We have a significant number of customers ‘aged’ 18 in the dataset, however it is due to the default age in the system. We do not know the real age of these customers.

They are excluded for the following statistical tests.

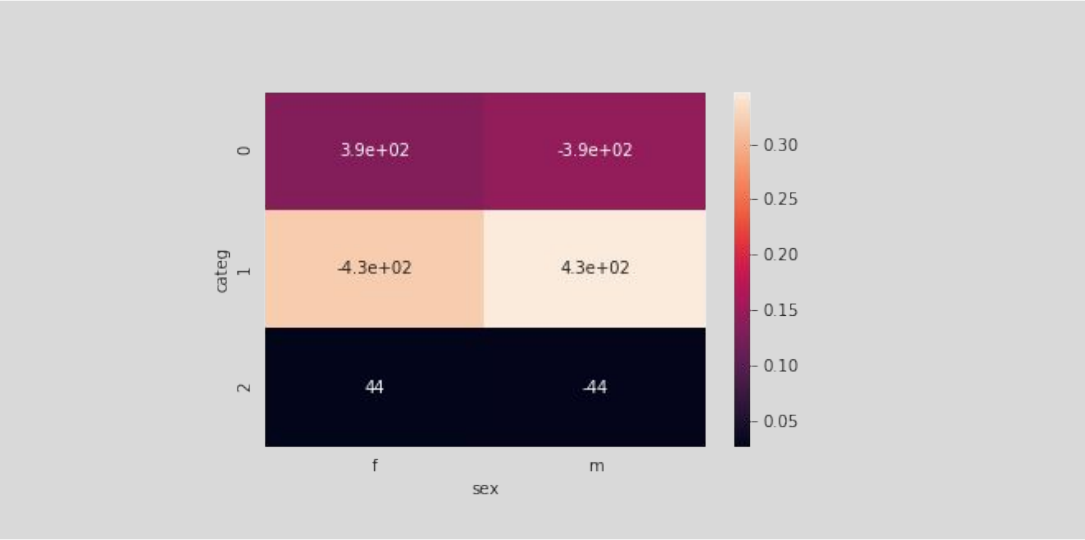
Preliminary note about assumptions

There are 3 assumptions that need to be met for the results of an ANOVA test to be considered accurate and trust-worthy. Assumptions apply to the residuals, not the variables.

- Normality
- Homogeneity of variance (homoscedasticity)
- Independent observations

However due to the number of observations (much larger than 50) in each group, we can accept the results of the ANOVA.

Men are buying more products from category 1



Bivariate Analysis

Hypothesis with **Chi2-test** :

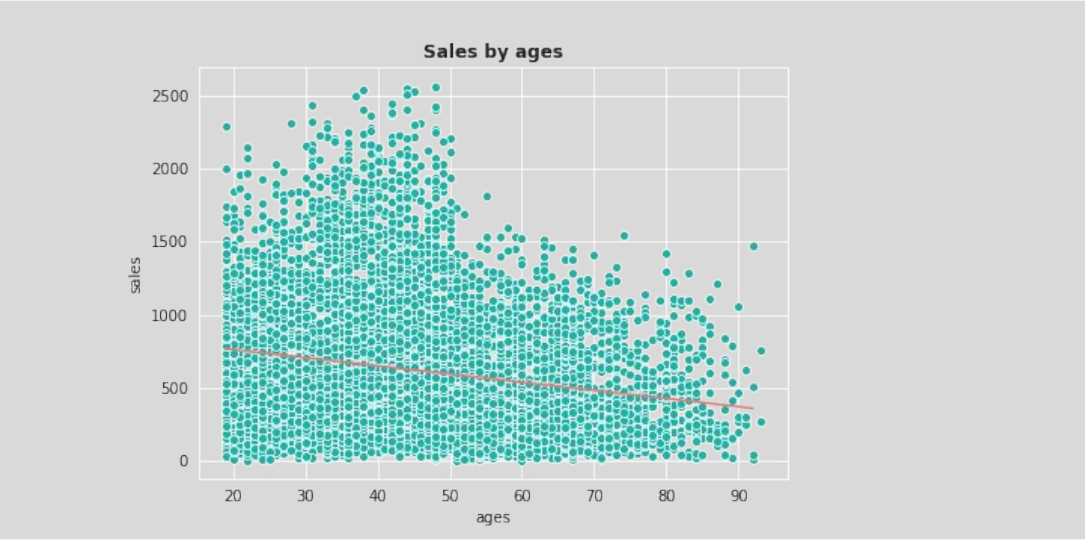
- H0: There is no correlation between gender and category
- H1: There is a correlation between gender and category

Chi-2: 11.19, Degree of freedom: 2

p-value : **0.003709**

We can reject H0, there is a **correlation between gender and category**.

Bivariate Bravais-Pearson Correlation not conclusive



Bivariate Analysis - Bravais-Pearson

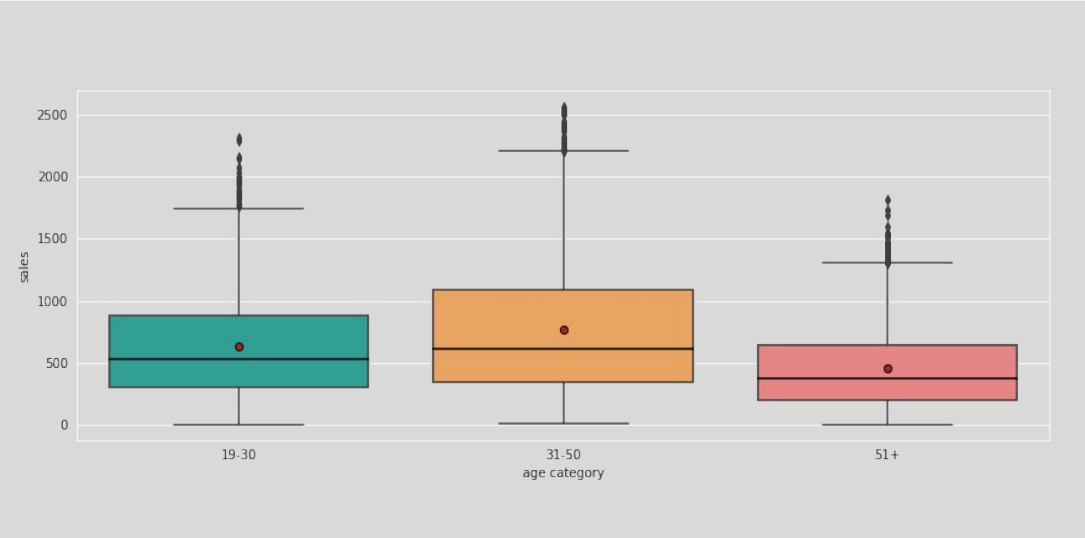
Hypothesis :

- H0: There is no correlation between ages and sales
- H1: There is a correlation between ages and sales

R-Squared: **0.039,** p-value : **3.53e-73**

We can reject H0 however the interpretation of these results is limited.

Customers from '31-50' are spending more



Bivariate Analysis - ANOVA

Hypothesis :

- H0: There is no correlation between age categories and sales
- H1: There is a correlation between age categories and sales

R-Squared: **0.089,** p-value : **1.89e-166**

We can reject H0 (p-value < 0.05), there is a **medium correlation** between age categories and sales.

Age categories explained **8.9%** of the variation in sales.

Bivariate Bravais-Pearson Correlation not conclusive

Bivariate Analysis - Bravais-Pearson

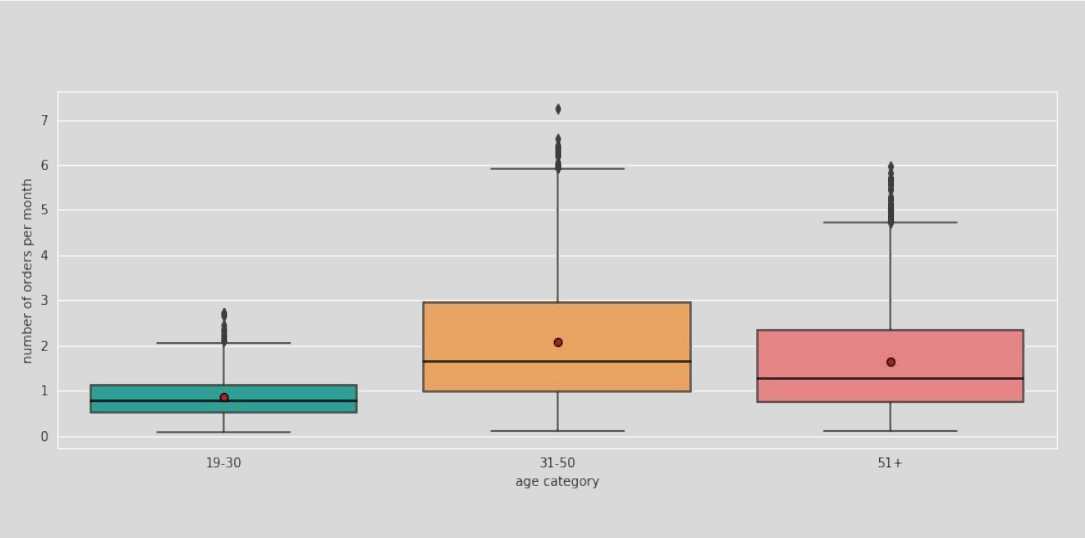
Hypothesis :

- H0: There is no correlation between ages and number of orders per month
- H1: There is a correlation between ages and number of orders per month

R-Squared: **0.017,** p-value : **5.72e-32**

We can reject H0 however the interpretation of these results is limited.

Customers from '31-50' are ordering more often



Bivariate Analysis - ANOVA

Hypothesis :

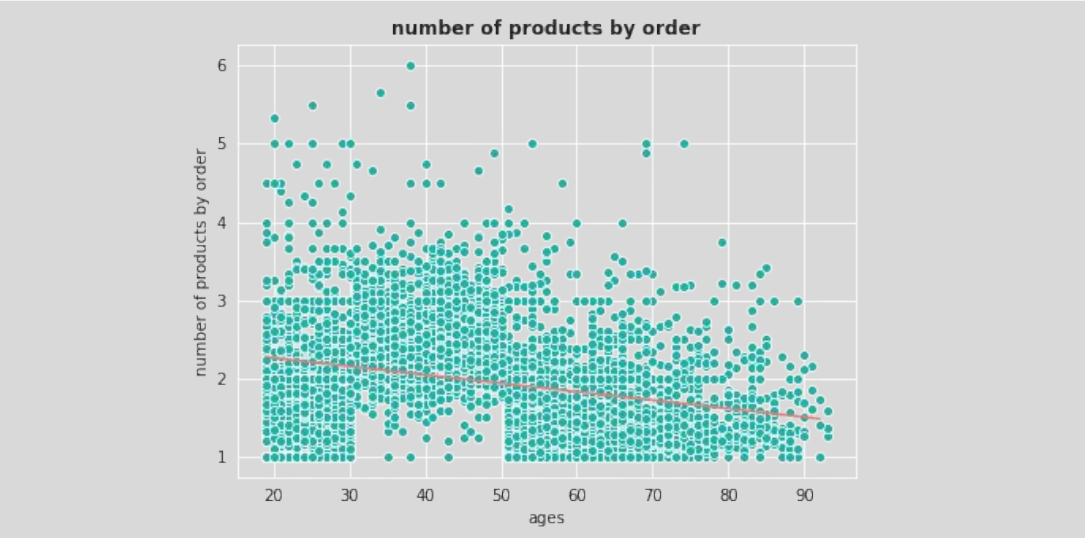
- H0: No correlation between number of orders per month and age categories
- H1: Correlation between number of orders per month and age categories

R-Squared: **0.137,** p-value : **4.39e-262**

We can reject H0 (p-value < 0.05), there is a **medium correlation** between age categories and number of orders per month.

Age categories explained **13.7%** of the variation in the number of orders per month.

Bivariate Bravais-Pearson Correlation not conclusive



Bivariate Analysis - Bravais-Pearson

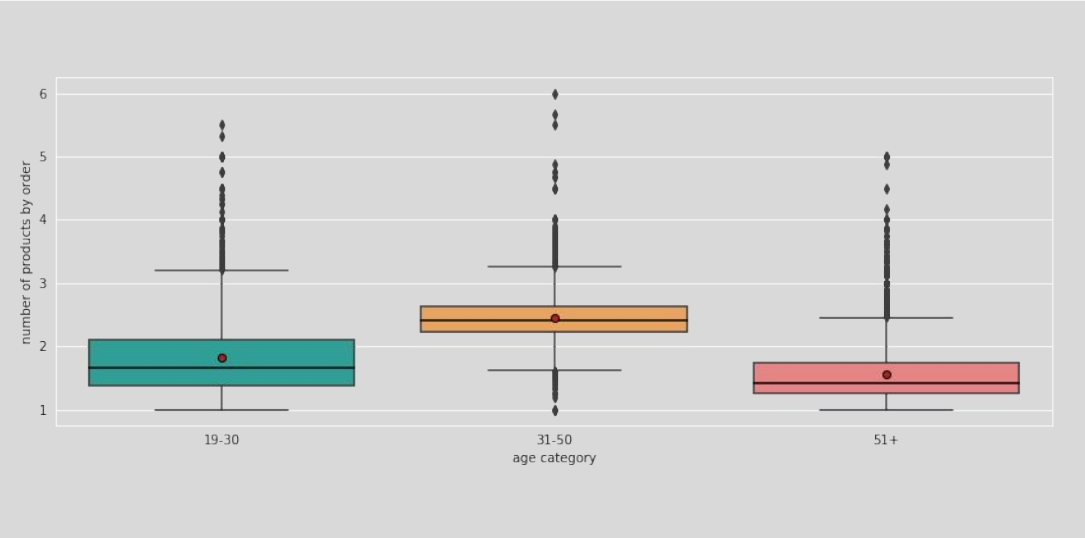
Hypothesis :

- H0: There is no correlation between ages and number of products per order
- H1: There is a correlation between ages and number of products per order

R-Squared: **0.075,** p-value : **2.41e-140**

We can reject H0 however the interpretation of these results is limited.

Customers from '31-50' are purchasing more products per order



Bivariate Analysis - ANOVA

Hypothesis :

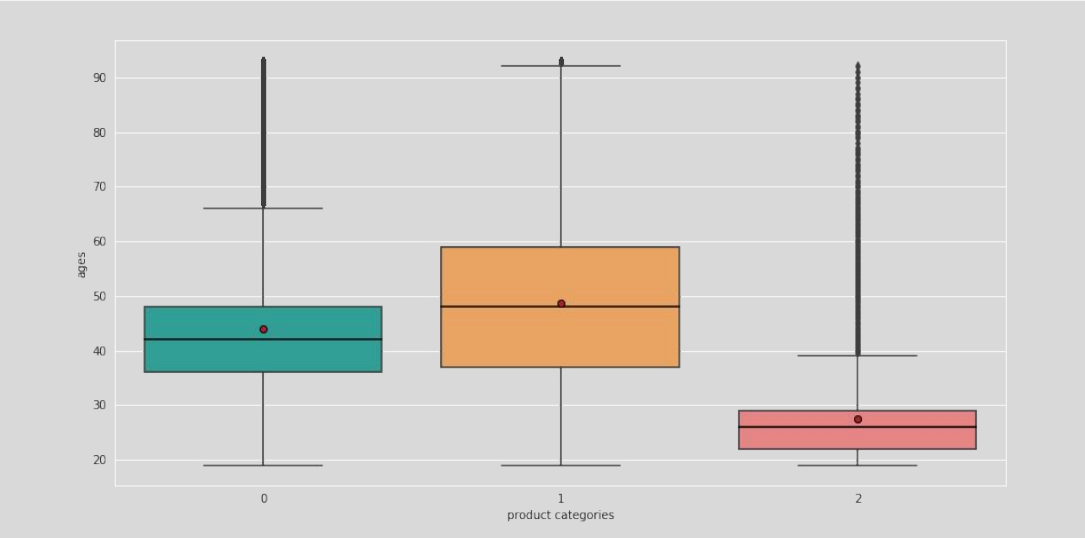
- H0: No correlation between nb of products by order and age categories
- H1: Correlation between nb of products by order and age categories

R-Squared: **0.394,** p-value : **0.00**

We can reject H0 (p-value < 0.05), there is a **strong correlation** between age categories and number of orders per month.

Age categories explained **39.40%** of the variation in the number of products per order.

Category 2 products are purchased by younger customers



Bivariate Analysis - ANOVA

Hypothesis :

- H0: No correlation between customer ages and product categories
- H1: Correlation between customer ages and product categories

R-Squared: **0.099,** p-value : **0.00**

We can reject H0 (p-value < 0.05), there is a **medium correlation** between product categories and customer ages.

Product categories explained **9.9%** of the variation in the customer ages.

THANK YOU!

Do you have any questions?

CREDITS: This presentation template was inspired from **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Pixabay**