SALES ANALYSIS

A close look to Captain book latest sales performance and insights.

INITIAL DATA ANALYSIS EXPLORATION.

- How the data set looks likes after clean from duplicates and outliers?
- What are the indicators of central tendency and dispersion.
- Visualising some interesting distribution and indicators.
- A concentration analysis, using a Lorenz curve and a Gini coefficient.
- Prize range distribution

	ID_PROD	DATE	SESSION_ID	CLIENT_ID
34387	T_0	test_2021-03-01 02:30:02.237443	s_0	ct_0
54813	T_0	test_2021-03-01 02:30:02.237412	s_0	ct_1
57261	T_0	test_2021-03-01 02:30:02.237439	s_0	ct_1
58802	T_0	test_2021-03-01 02:30:02.237429	s_0	ct_0
60170	T_0	test_2021-03-01 02:30:02.237446	s_0	ct_0

FOUND OF DUPLICATES AND TEST DATA

During the first analysis of the upload date set, I had observer that there was some duplicates values on the transaction dataset this had being test data.

	ID_PROD	DATE	SESSION_ID	CLIENT_ID	SEX	BIRTH	PRICE	CATEG	AGE
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977	4.99	0	44
1	0_1483	2021-12-27 11:11:12.123067	s_140787	c_5433	f	1981	4.99	0	40
2	0_1483	2021-10-27 04:56:38.293970	s_110736	c_857	m	1985	4.99	0	36
3	0_1483	2021-07-04 06:43:45.676567	s_57626	c_3679	f	1989	4.99	0	32
4	0_1483	2021-09-19 08:45:43.735331	s_92165	c_1609	m	1980	4.99	0	41

HOW THE DATA SET LOOKS LIKES AFTER CLEAN FROM DUPLICATES AND OUTLIERS? THE FIRST 5 ROWS

After cleaning the all data set I could create a new data set merge from them and use it for further analysis. This step is done with the TL script available on the project it exports the data_clean.zip file containing the data.csv imported

	ID_PROD	DATE	SESSION_ID	CLIENT_ID	SEX	BIRTH	PRICE	CATEG	AGE
336708	0_1920	2021-04-13 18:36:10.252971	s_20115	c_7088	m	1987	25.16	0	34
336709	0_1920	2021-05-30 02:37:22.371278	s_41465	c_7748	f	1989	25.16	0	32
336710	2_23	2021-09-27 04:47:02.271354	s_96170	c_3976	f	1992	115.99	2	29
336711	2_28	2021-05-11 01:31:34.932056	s_32812	c_7613	f	1993	103.5	2	28
336712	2_98	2021-03-08 21:10:32.250919	s_3637	c_5967	f	2003	149.74	2	18

THE LAST 5 ROWS

	BIRTH	PRICE	CATEG	AGE
COUNT	336713.0	336713.0	336713.0	336713.0
MEAN	1977.8235678456133	17.21518851366	0.43015565184593 41	43.176432154386674
STD	13.524433308700898	17.855445377654487	0.5910818586413054	13.524433308700898
MIN	1929.0	0.62	0.0	17.0
25 %	1971.0	8.61	0.0	34.0
50 %	1980.0	13.9	0.0	41.0
75 %	1987.0	18.99	1.0	50.0
MAX	2004.0	300.0	2.0	92.0

THE TABLE SHOW INFORMATION ABOUT ITS MAIN INDICATORS IN PARTICULAR **BIRTH AND PRICE**

- Minimum year value 1929 maximum year value 2004.
- There are 75 years values observed.
- Price Mean is 17.215189.
- Price min 0.62 and max 300.00

	MEDIAN	SKEW	KURTOSIS	VAR
BIRTH	1980.0	-0.5804173699368993	0.4523350716075605	182.9102963214983
PRICE	13.9	5.479196379351584	45.42520484470874	318.81692963440304
CATEG	0.0	1.0288559773127675	0.0496156969356778	0.34937776361486006
AGE	41.0	0.5804173699369055	0.4523350716075614	182.9102963214983

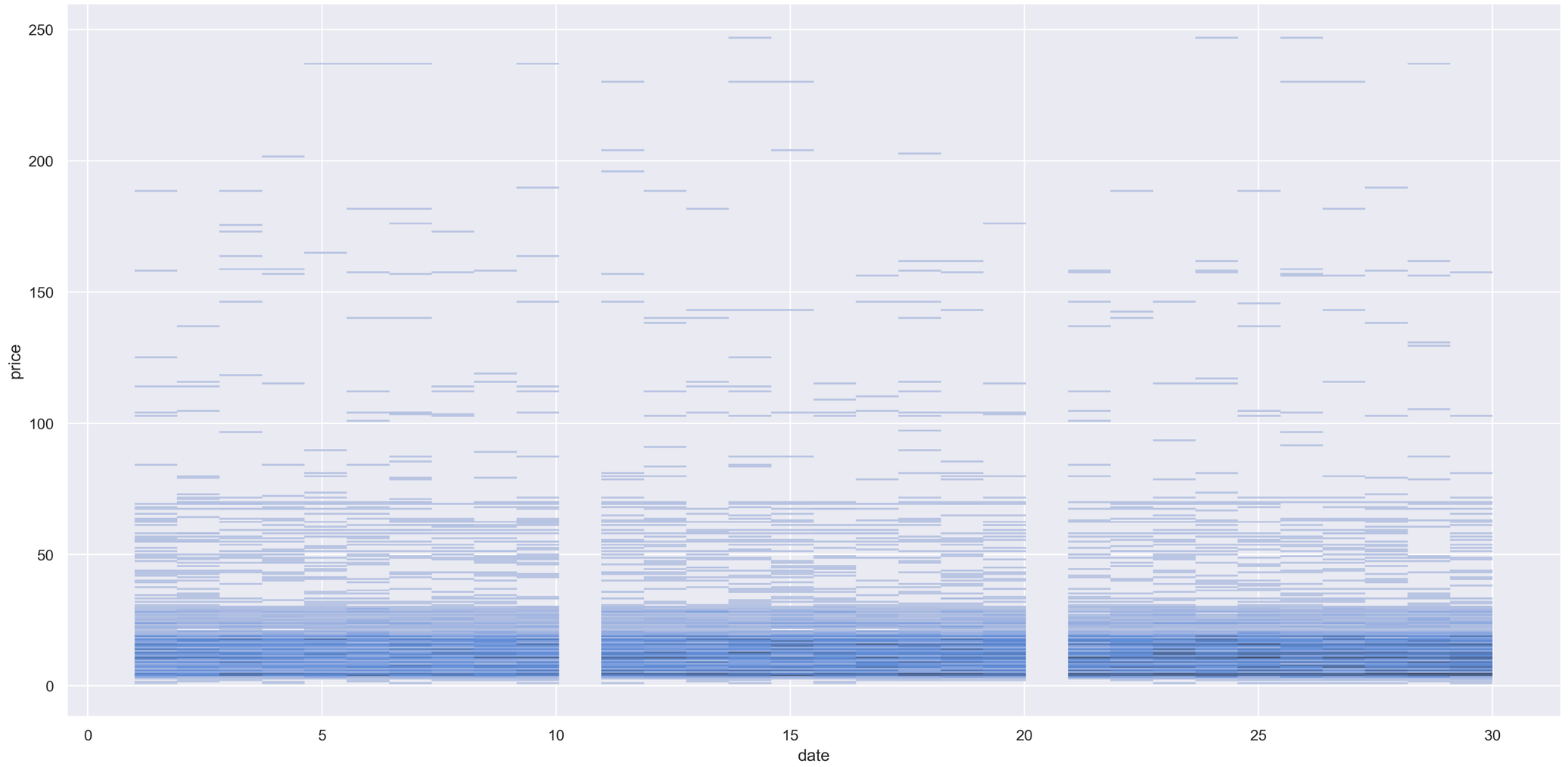
INDICATORS OF CENTRAL TENDENCY AND DISPERSION

This are some indicators that tell us about the variance and symmetry of this data set.

	BIRTH	PRICE	CATEG	AGE
COUNT	21577.000000	21577.000000	21577.000000	21577.000000
MEAN	1978.866710	14.791674	0.184734	42.133290
STD	12.243654	17.539109	0.508076	12.243654
MIN	1929.000000	0.620000	0.000000	17.000000

MONTH DISTRIBUTION ANOMALY ANALYSIS

Describing October and getting info.
It has a very low minimal price but it not the cause of its anomaly. It has also fewer entries and therefore less memory usage as we can see in the next slide.

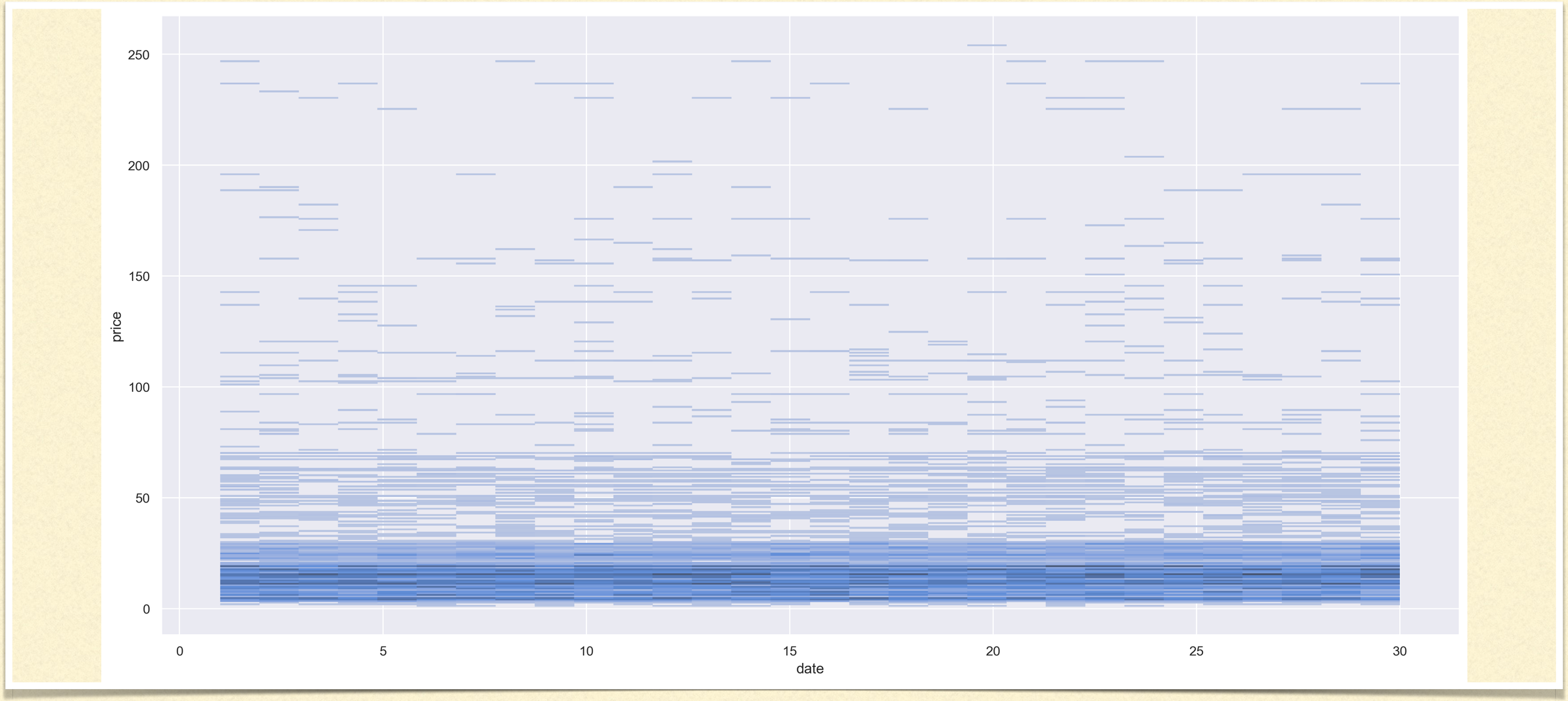


VISUALISING SEPTEMBER PRICES AND DATE.

Check for anomaly in months day prices values. September looks regular.

VISUALISING OCTOBER PRICES AND DATE.

Check for anomaly in months day prices values. October has clearly an anomaly here!



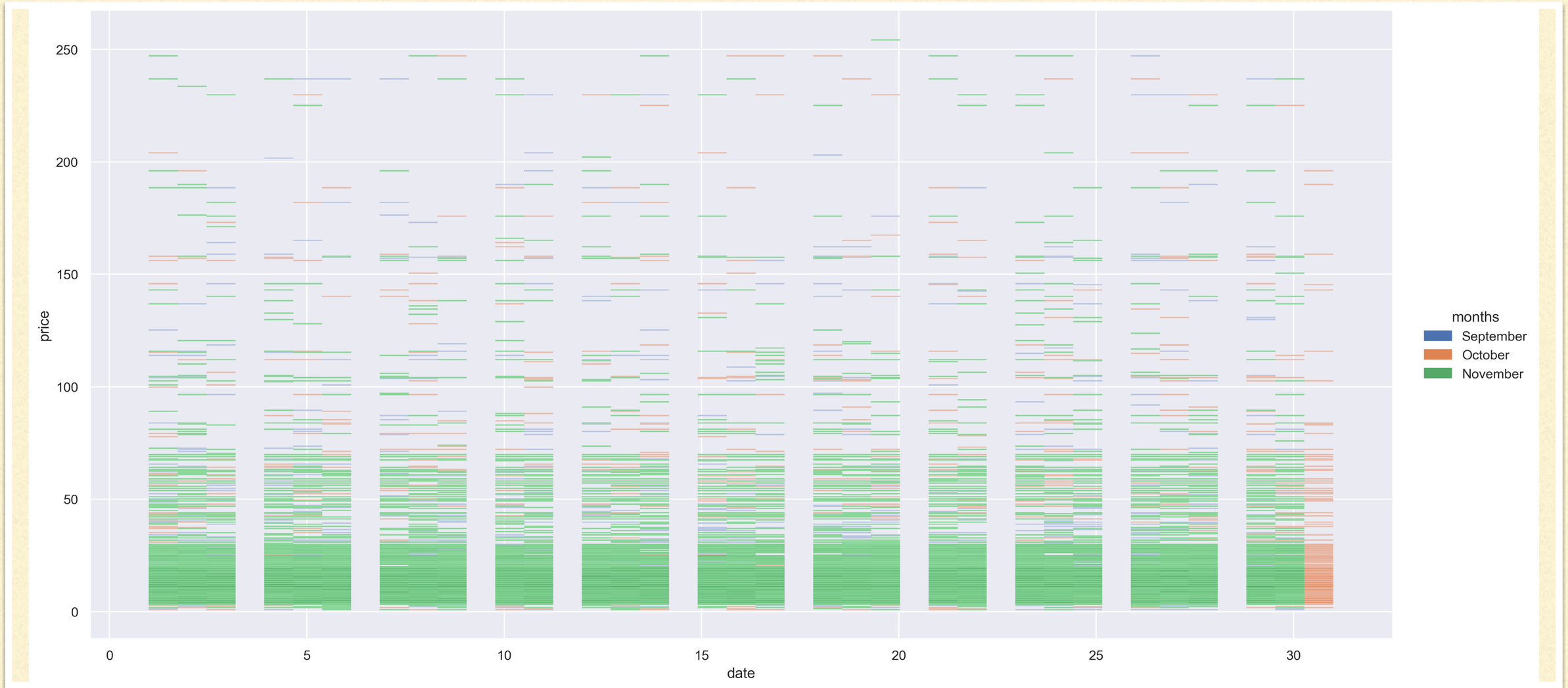
VISUALISING NOVEMBER PRICES AND DATE.

Check for anomaly in months day prices values. November looks regular as September.

INFO	AUGUST	DECEMBER	FEBRUARY	JULY	NOVEMBER	OCTOBER	SEPTEMBER
N_OF_COLUMNS	10	10	10	10	10	10	10
N_OF_ROWS	25610	32417	29556	24712	28267	21577	33254
MEMORY	2.1+ MB	2.7+ MB	2.5+ MB	2.1+ MB	2.4+ MB	1.8+ MB	2.8+ MB

CONCLUSION

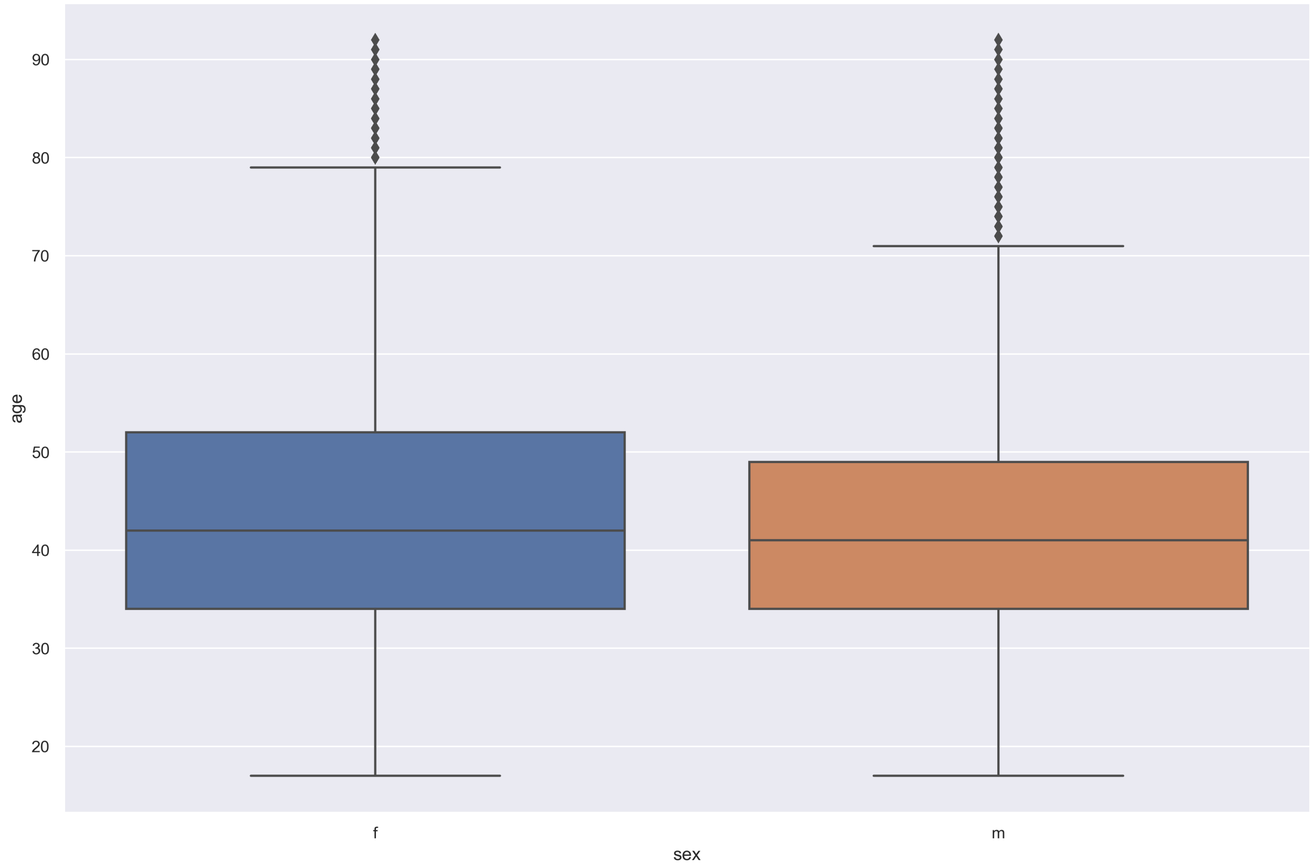
Conclusion: It just looks like that October has fewer rows then any others months and
that there is a significant amount of prices values between 20-40 from the start of the month to its end 30 being missing. Perhaps cause by a tech problem.



VISUALISING OCTOBER CLOSEST MONTHS.

This Visualisation show more clearly the missing prices values of October compare to September and November.

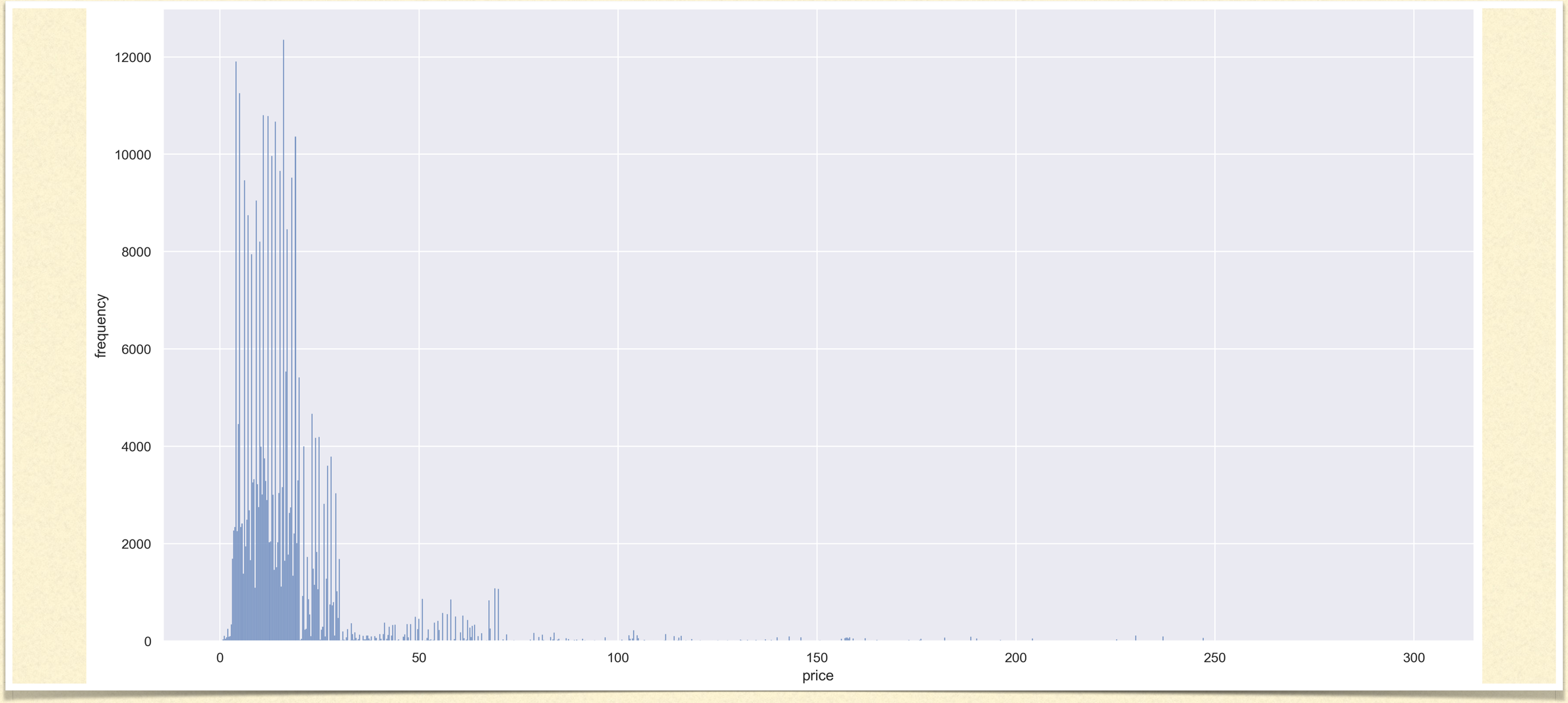
VISUALISING SOME INTERESTING DISTRIBUTIO N AND INDICATORS.



VISUALISING INDICATORS.

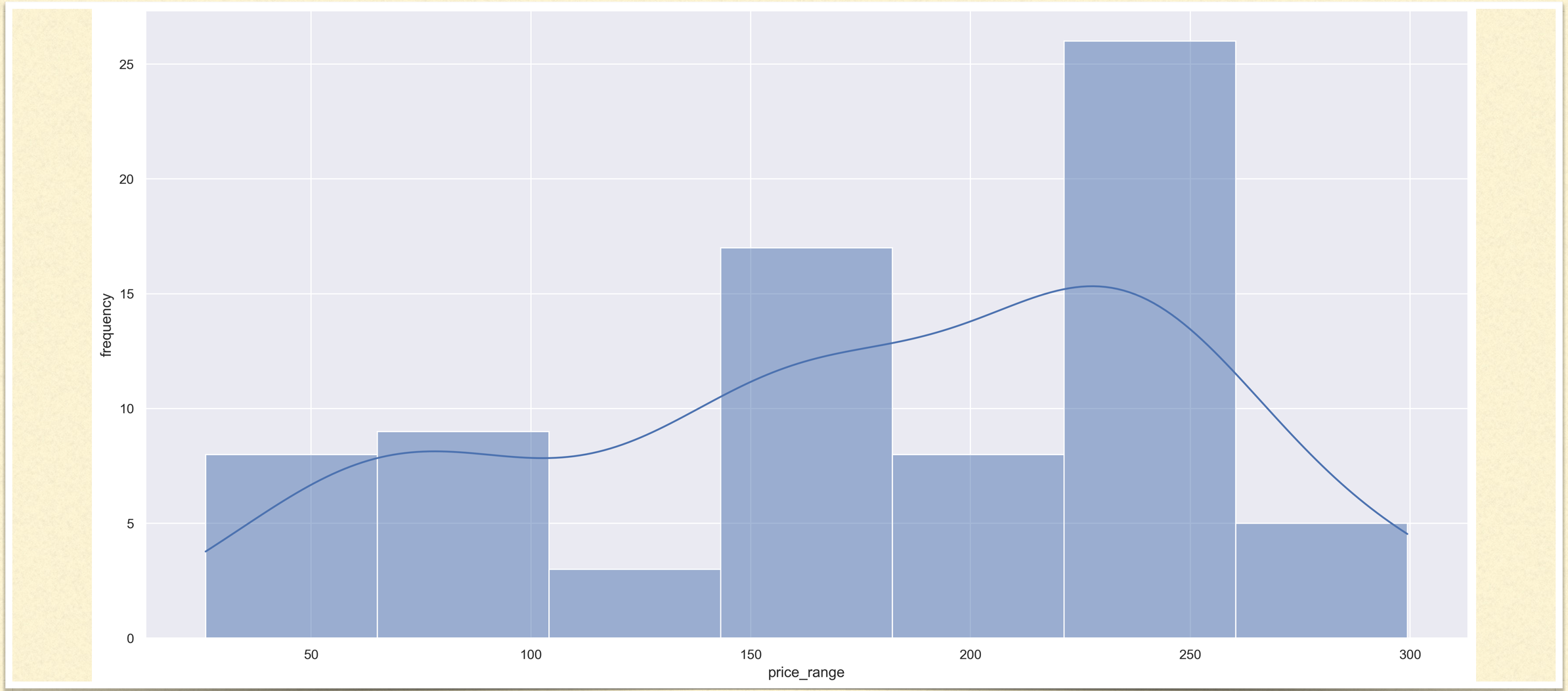
How is the distribution of age between gender? On males, they first quartile and the third are quite closer to they median comparing to female.

Female ones have fewer outliers compare to males. Both genders median are between 40-50s.



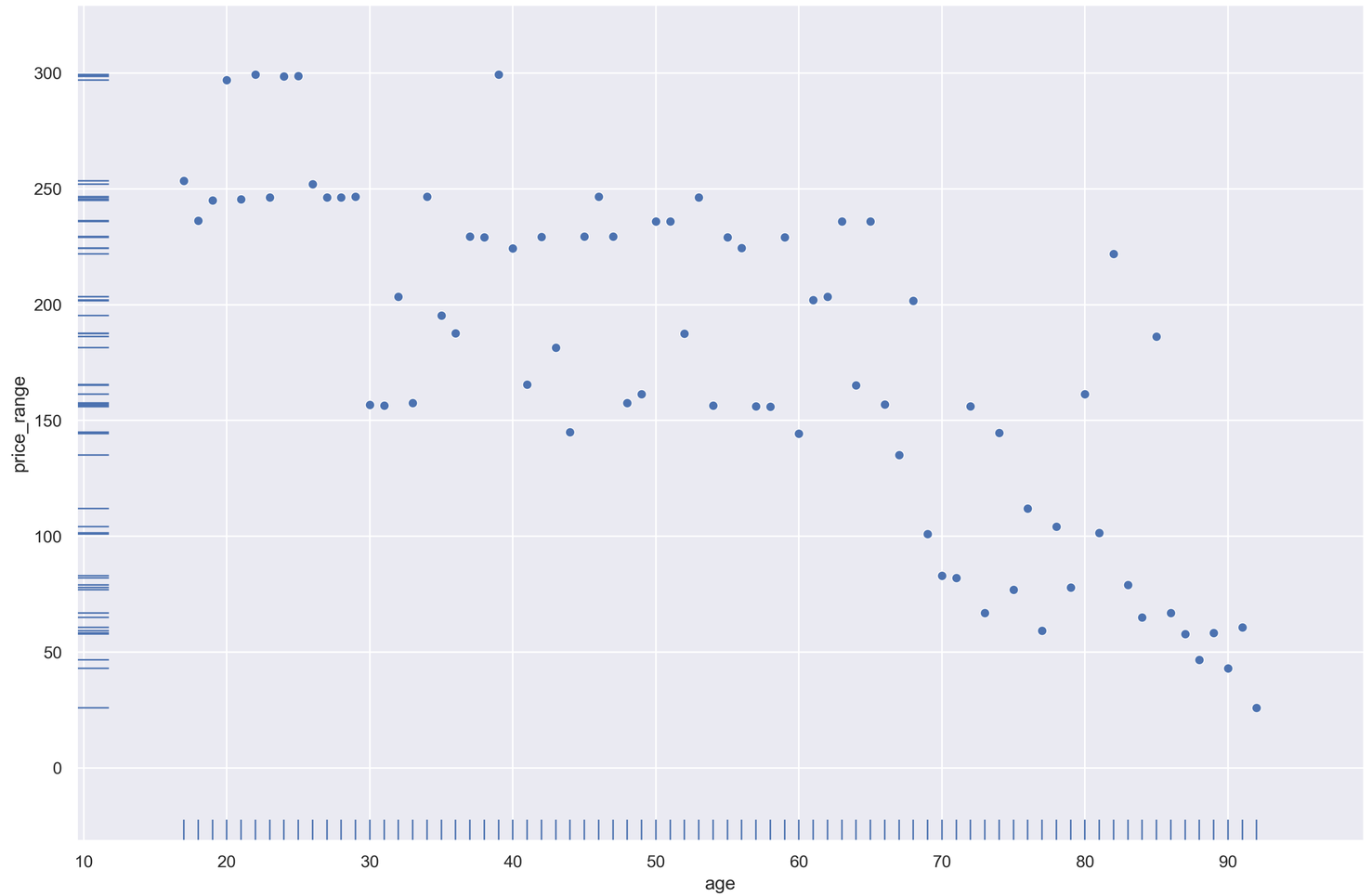
A POSITIVELY SKEWED PRICE.

This is because there is more low prices values on our data set. They are concentrated below the 100 price value.



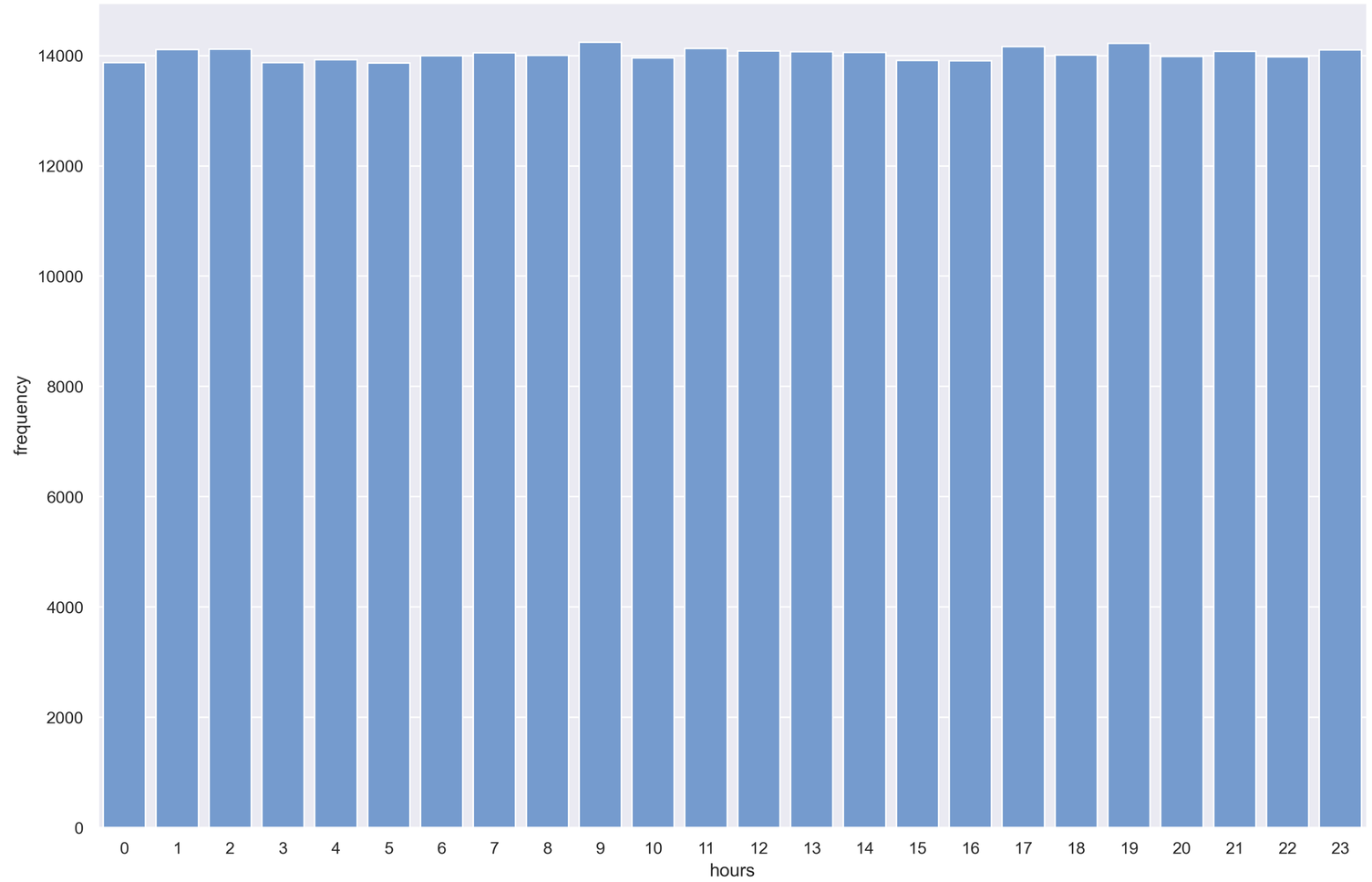
PRICE RANGE DISTRIBUTION

On this chart is clear to see where higher prices range are concentrated 150-250. its minimal 50 and max 300 have a difference of 250 with is minimal have almost a double of its frequency.



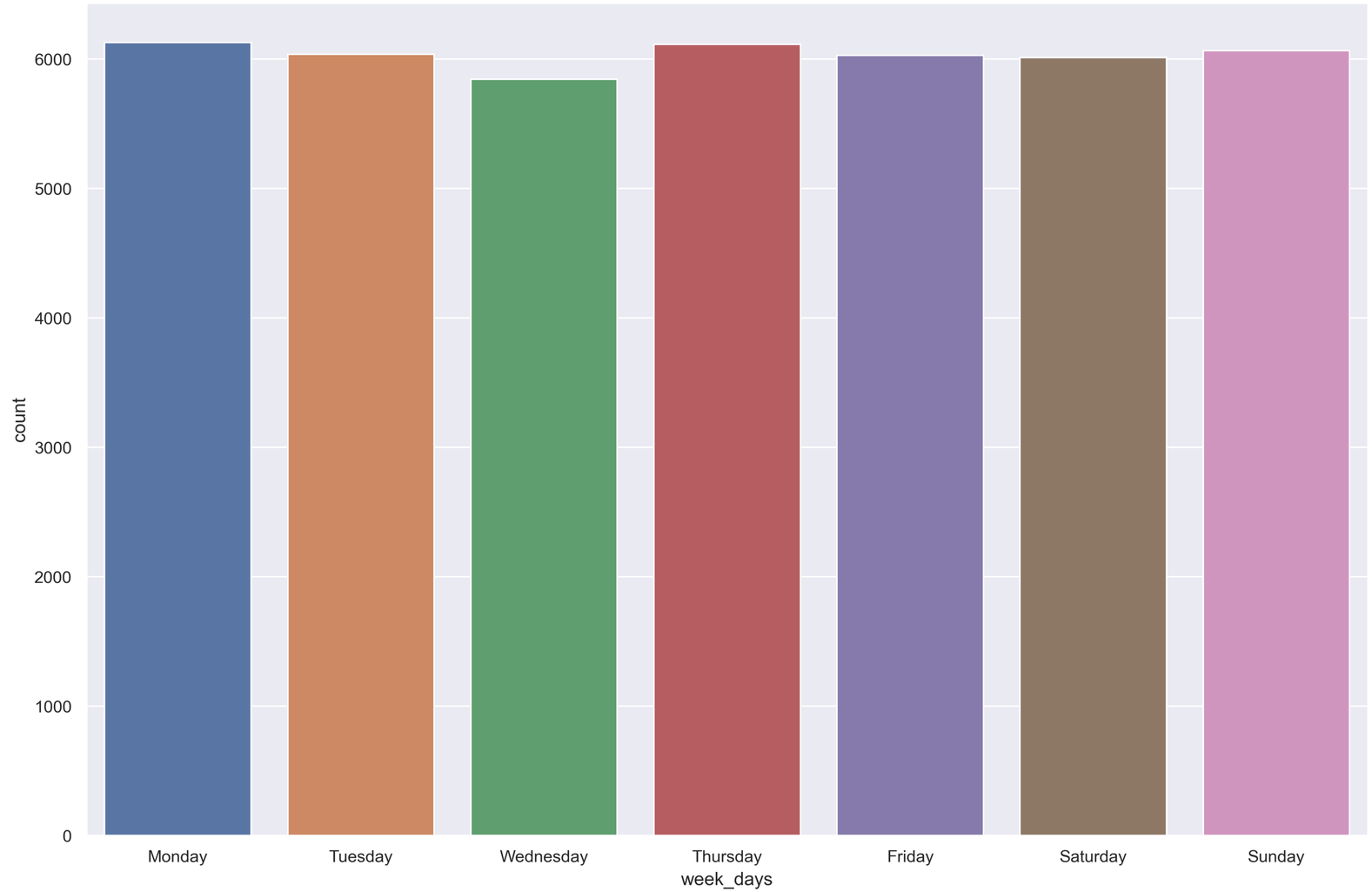
DISPERSION AND CONCENTRATION

Here is showing a more clear dispersion and concentration of price range and age. It seems that younger ones from 20-40 expend quite more then older ones 70-90.



TIME DISTRIBUTION

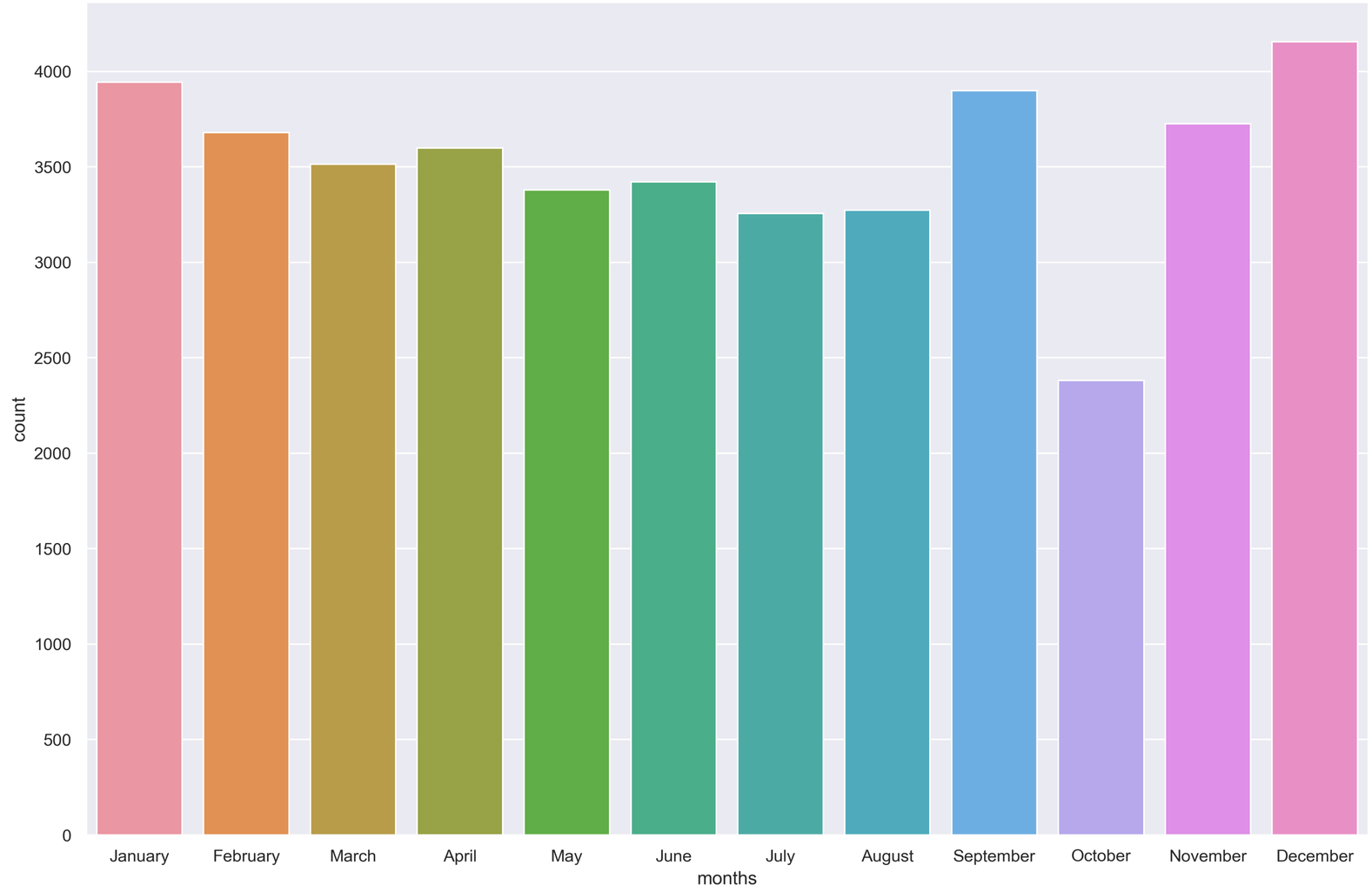
Here is the most frequent usage time. Is very homogenous, nevertheless there is a slight pick at 9 hour and 19 hour.



USERS SESSION COUNTS AND DISTRIBUTION

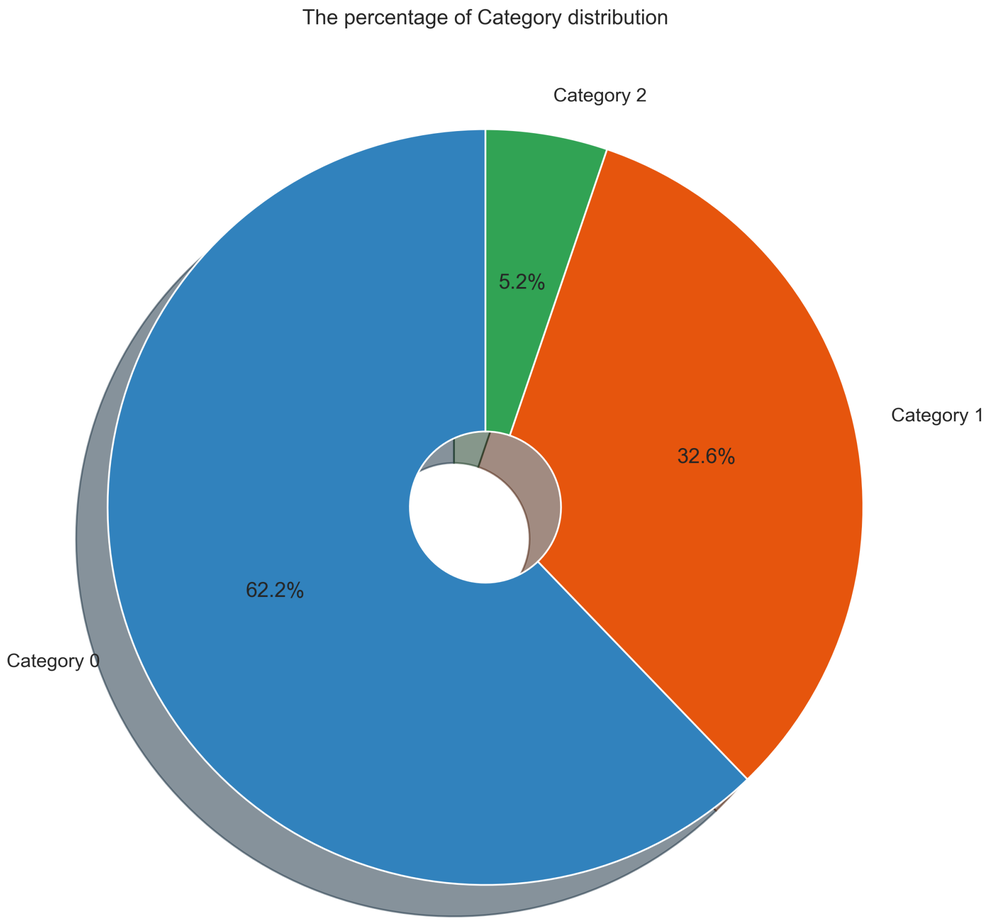
Based on the clients' id how is the amount of users distributed on the Weekdays?

The distribution is very flat here with small but notice higher values on Sunday, Monday and Thursday and quite low on Wednesday.



USERS SESSION MONTHLY DISTRIBUTION

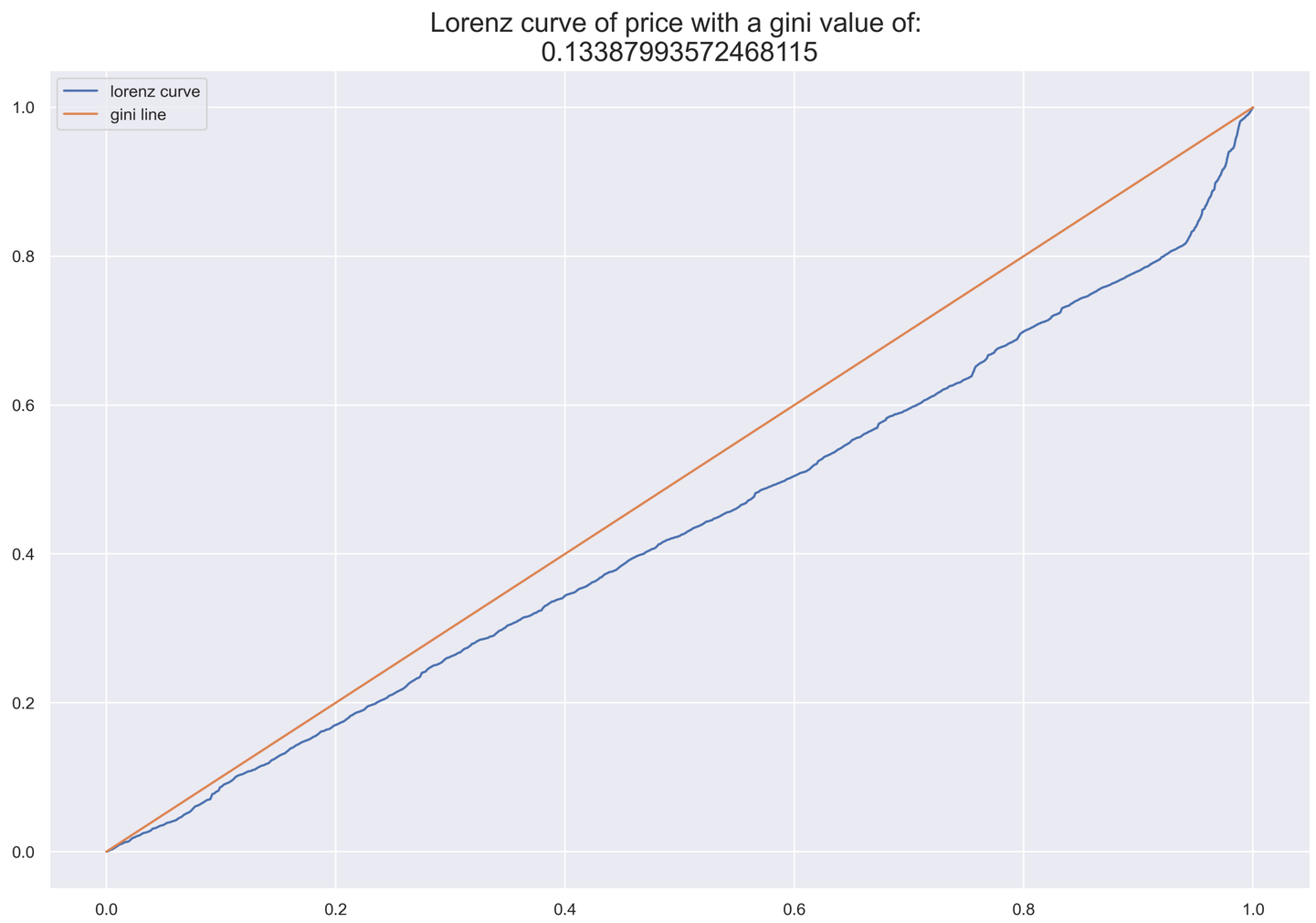
What then if we consider the amount of users per months how would the distribution look like? There is a quiet steadier decrease of users from February until August. This trend should be continued from September to October if there was not for a data anomaly October.



CATEGORY OF PRODUCTS DISTRIBUTION

How much percentage of each category?

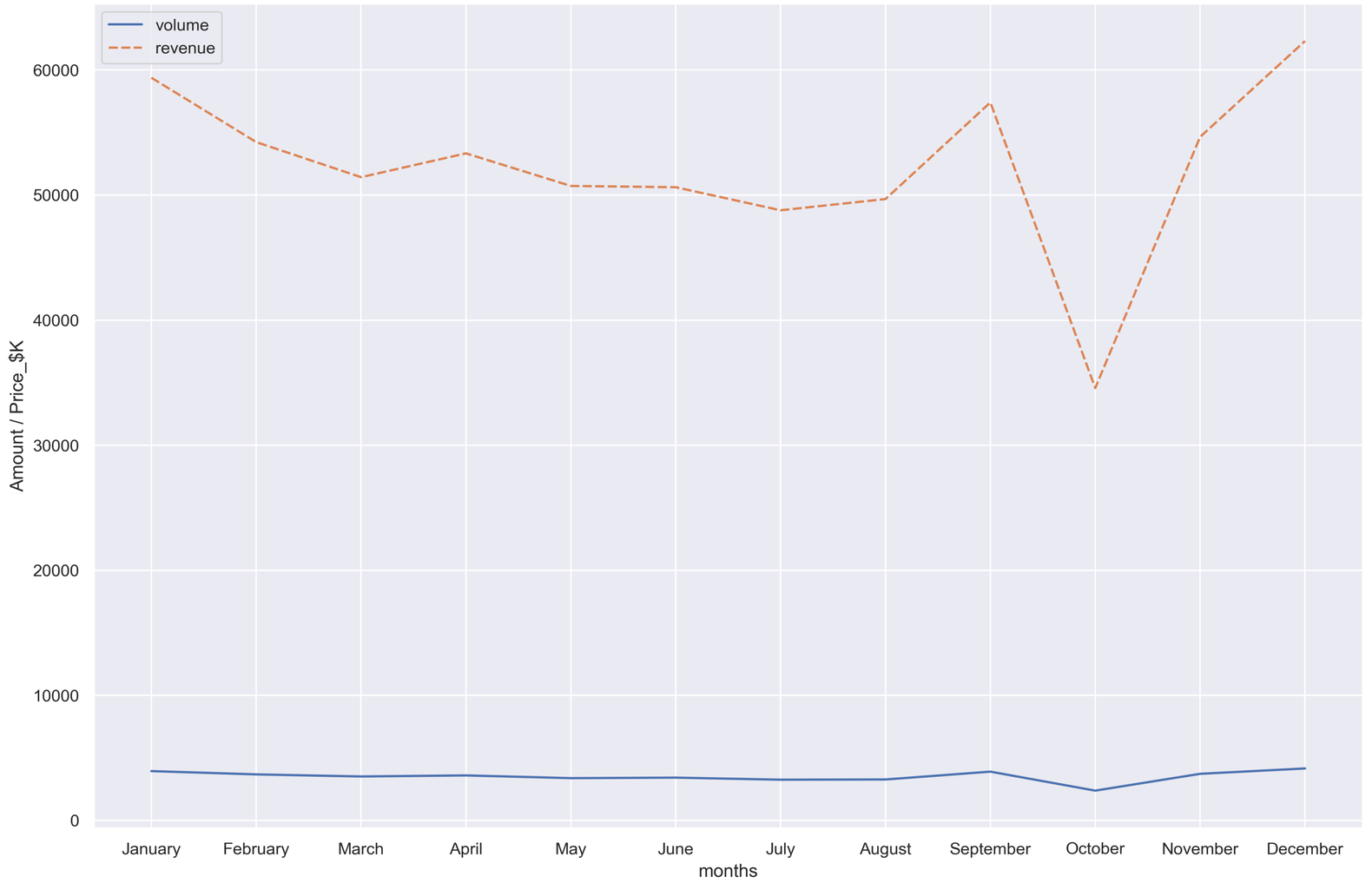
There is clear more observations of the category 0 and 1on the data set.



LORENZ CURVE AND GINI INDICATOR.

A concentration analysis, using a Lorenz curve and an its Gini coefficient indicator.

We can see clearing how price values are from lower to higher its look very uneven distributed.

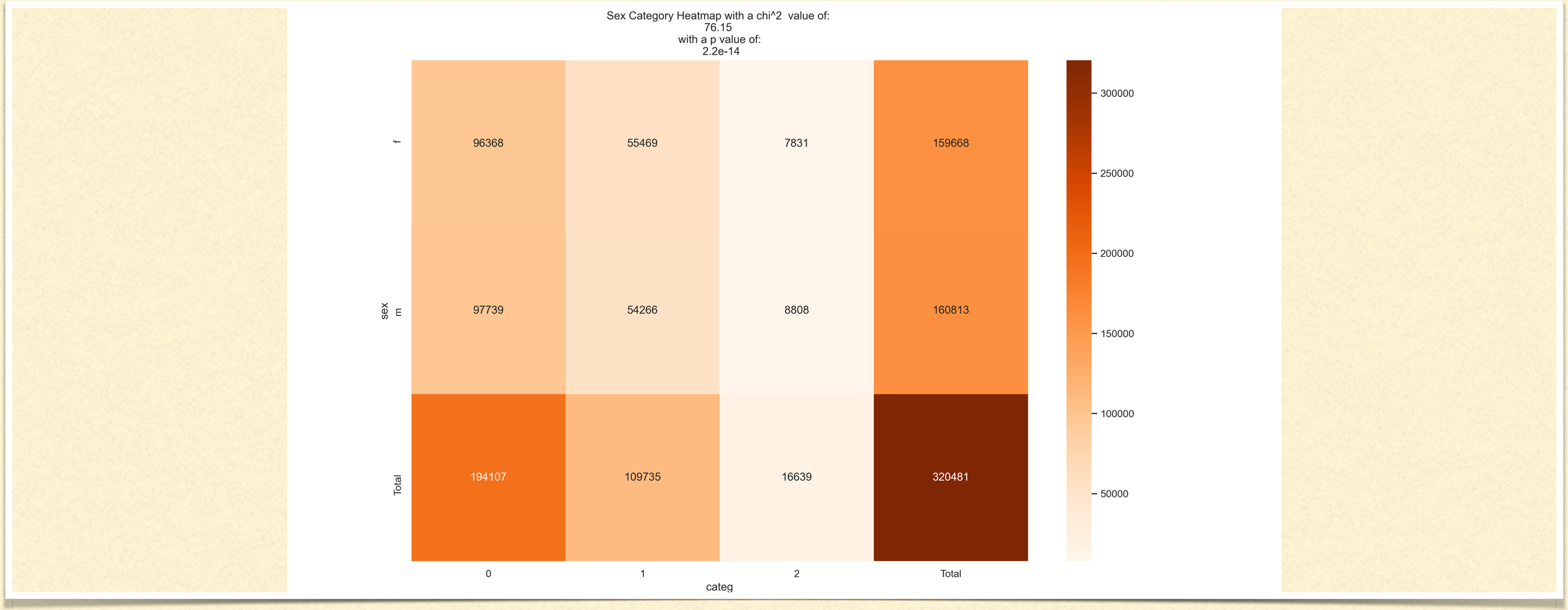


SALES VOLUME AND REVENUE.

This chart show the effect the lower users on the book store. The Volume and the Revenue fall on the already observed months October and July so as the is higher on November to February with the Highest at December.

SALES QUESTION AND ANALYSIS ANSWER

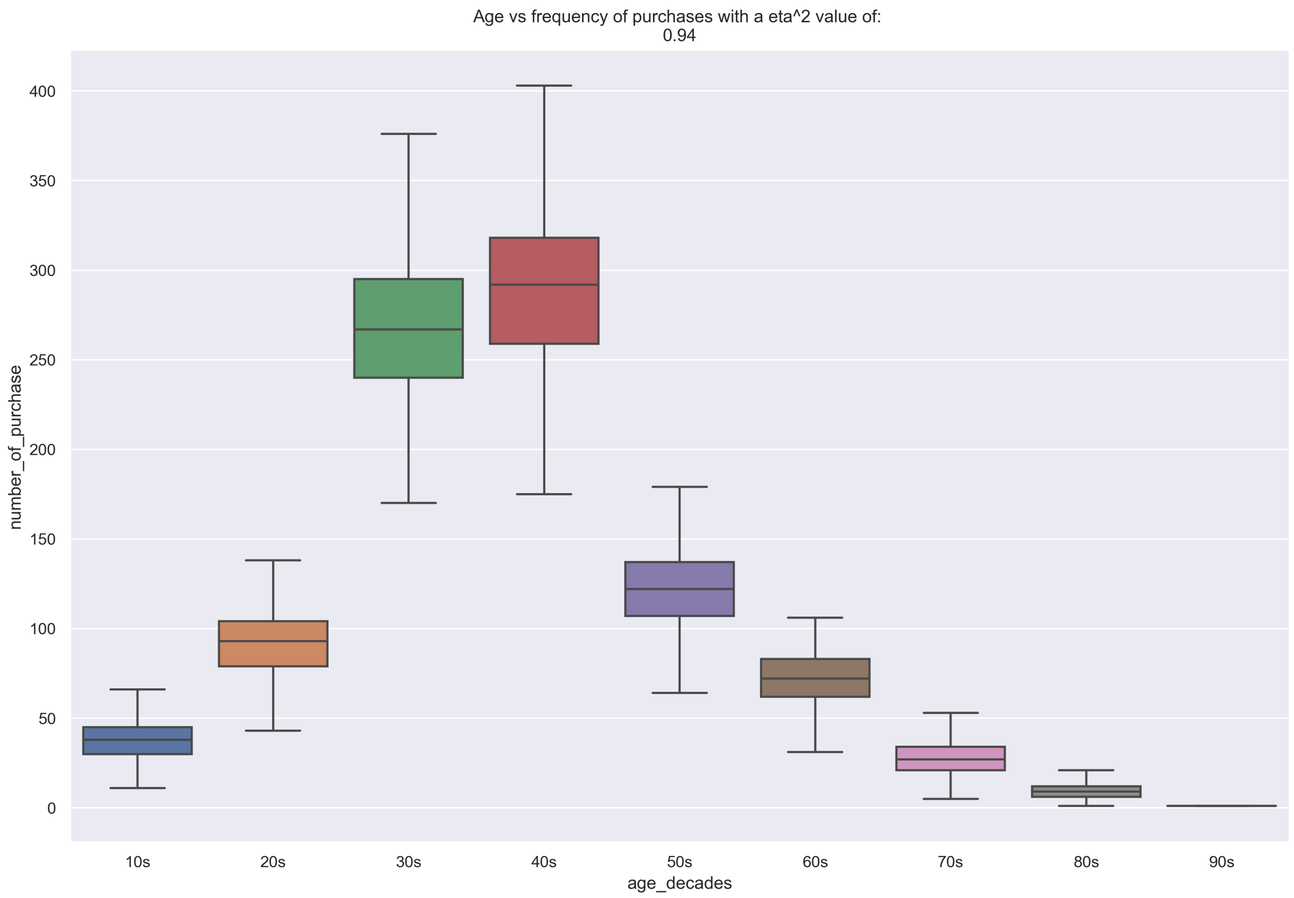
- Is there a correlation between gender and categories of products purchased?
- Is there a correlation between age and.
- The total amount of purchases?
- The purchase frequency (the number of purchases per month for example)?
- The average basket size (in number of items)?
- Categories of purchased products?



CORRELATION BETWEEN SEX AND CATEGORY.

Using Chi square test method this is the correlation between Sex a Categorical data and Category a Qualitative data.

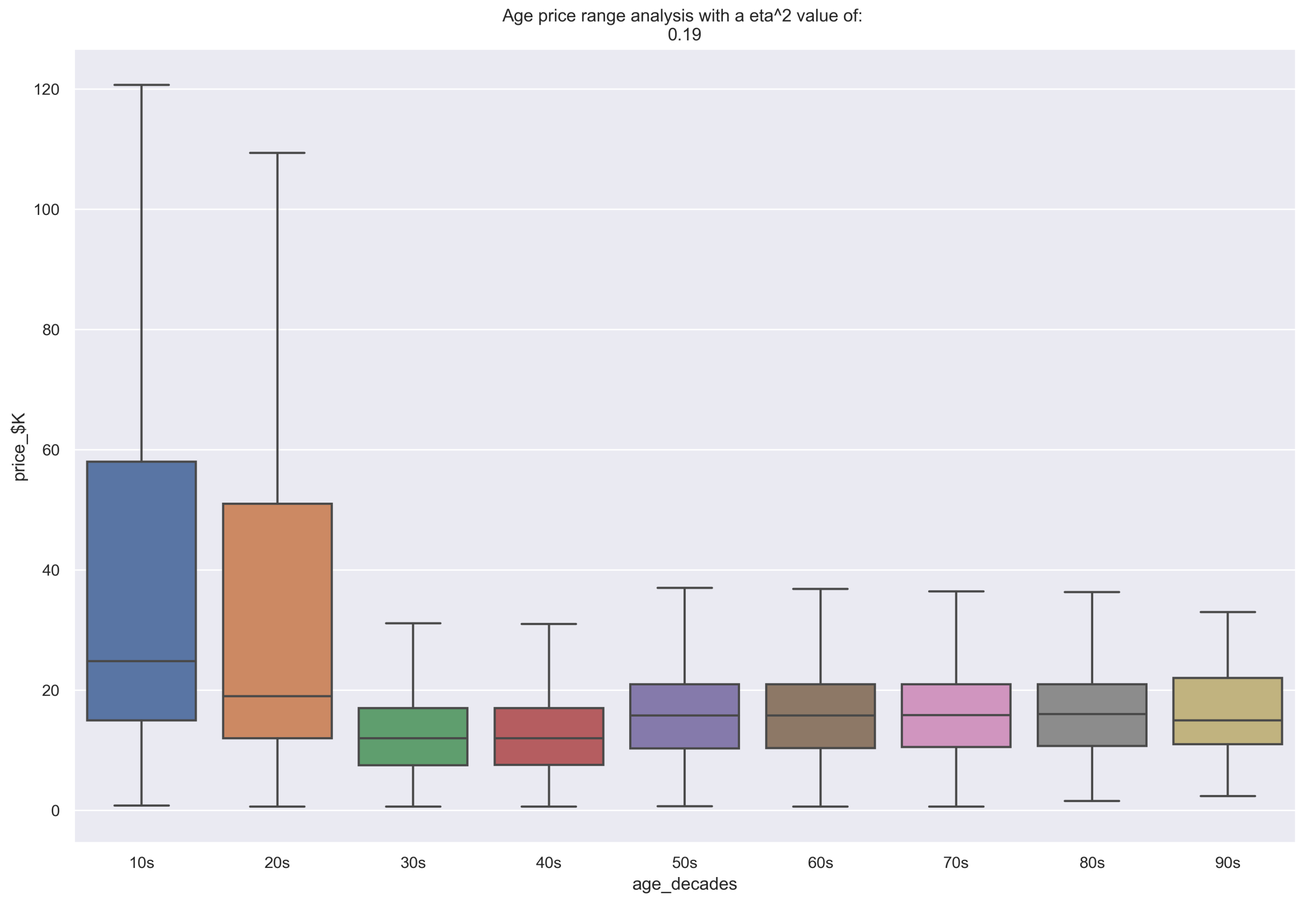
It Show have a Chi square of 75.15 and a p value of 2.2 so the null hypothesis can’t be reject.



AGE AND THE PURCHASES FREQUENCY.

Using the Eta square indicator of correlation between Quantitive variable age and price but considering the price frequency instead of its value.

There is a correlation value of 0.94 this indicate that age move together with the purchase amount. This will be more clean in nexts slides.

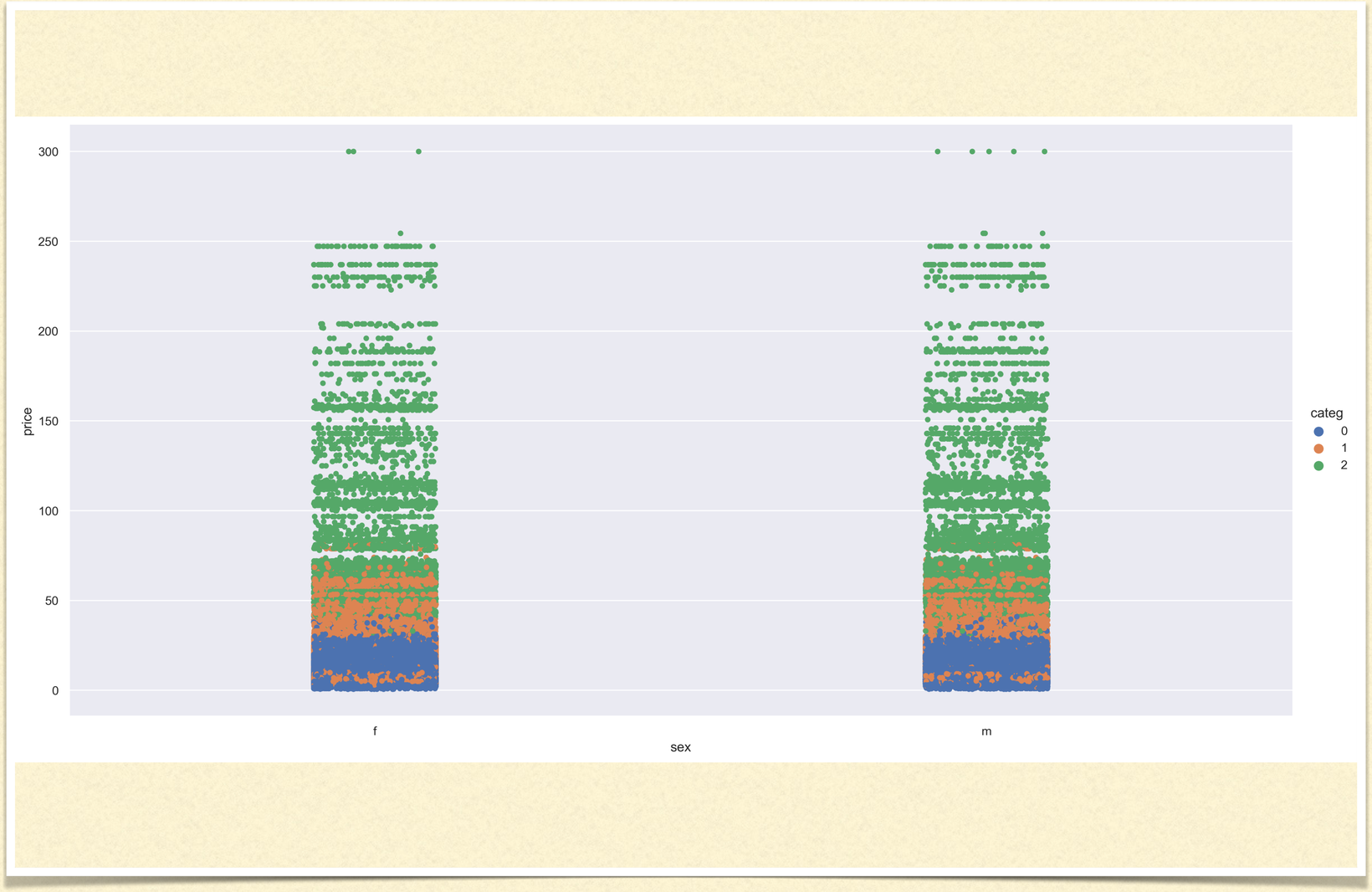


CORRELATION BETWEEN AGE AND PRICE RANGE?

Again using Eta square as the most appropriate correlation indicator, but this time considering the price range.

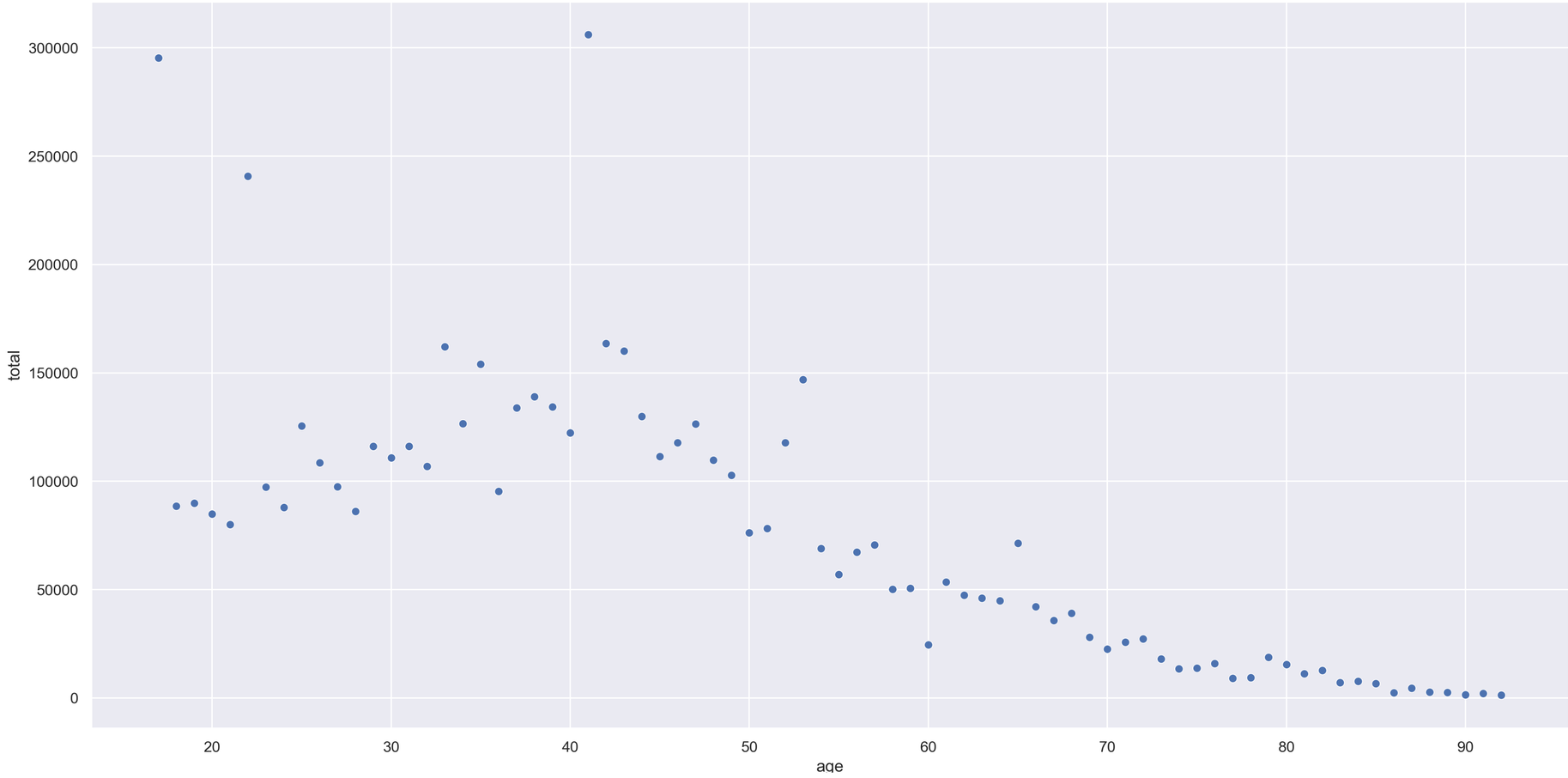
We get a different graphic here the influence of age is not so strong from 30s, 40s and so on. This is because Teenagers buy more expensive products.

IS THERE A CORRELATION BETWEEN GENDER AND CATEGORIES OF PRODUCTS PURCHASED?



Male seams too paid for more for costly products of the category 2 compare to female.

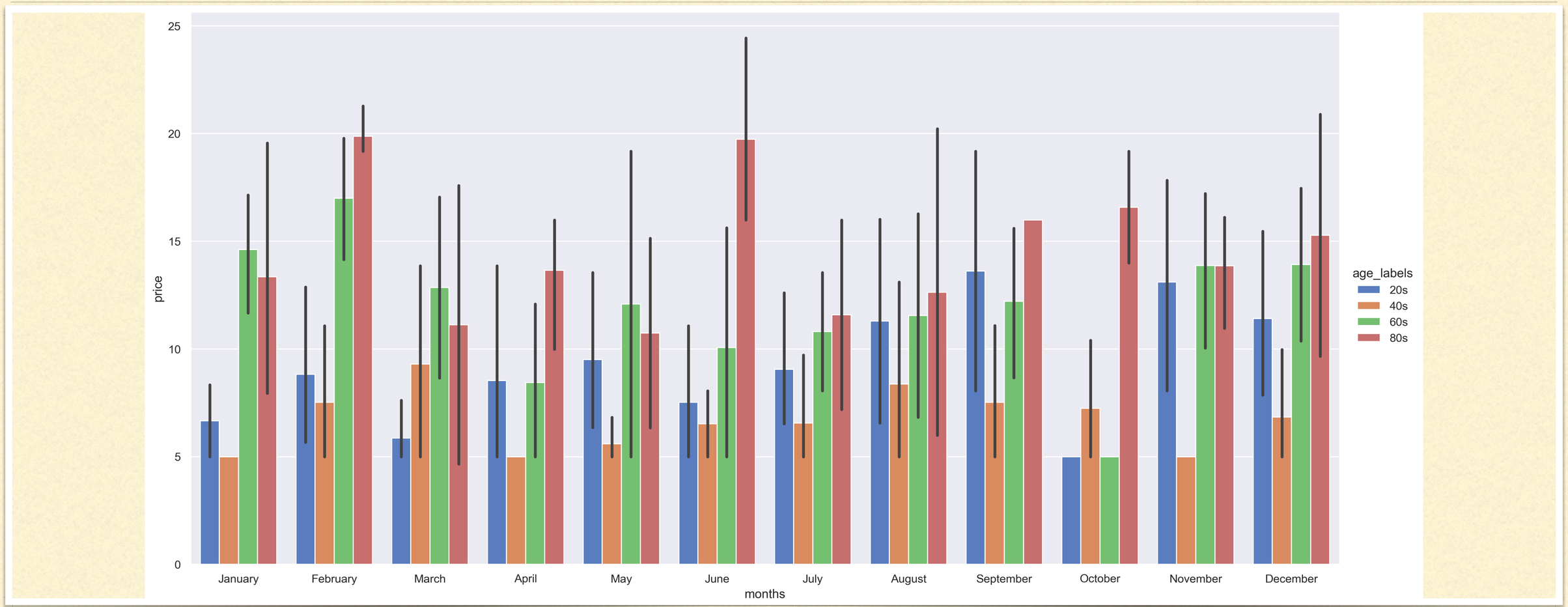
Lower values for category 1 and 2 are from bellow 100.



THE TOTAL AMOUNT OF PURCHASES?

As already spotted by the visualisation of sales, there is a clear difference in age and products purchased.

in the middle age 30 - 40 there are more costly products purchased but decline more the clients approach old age.

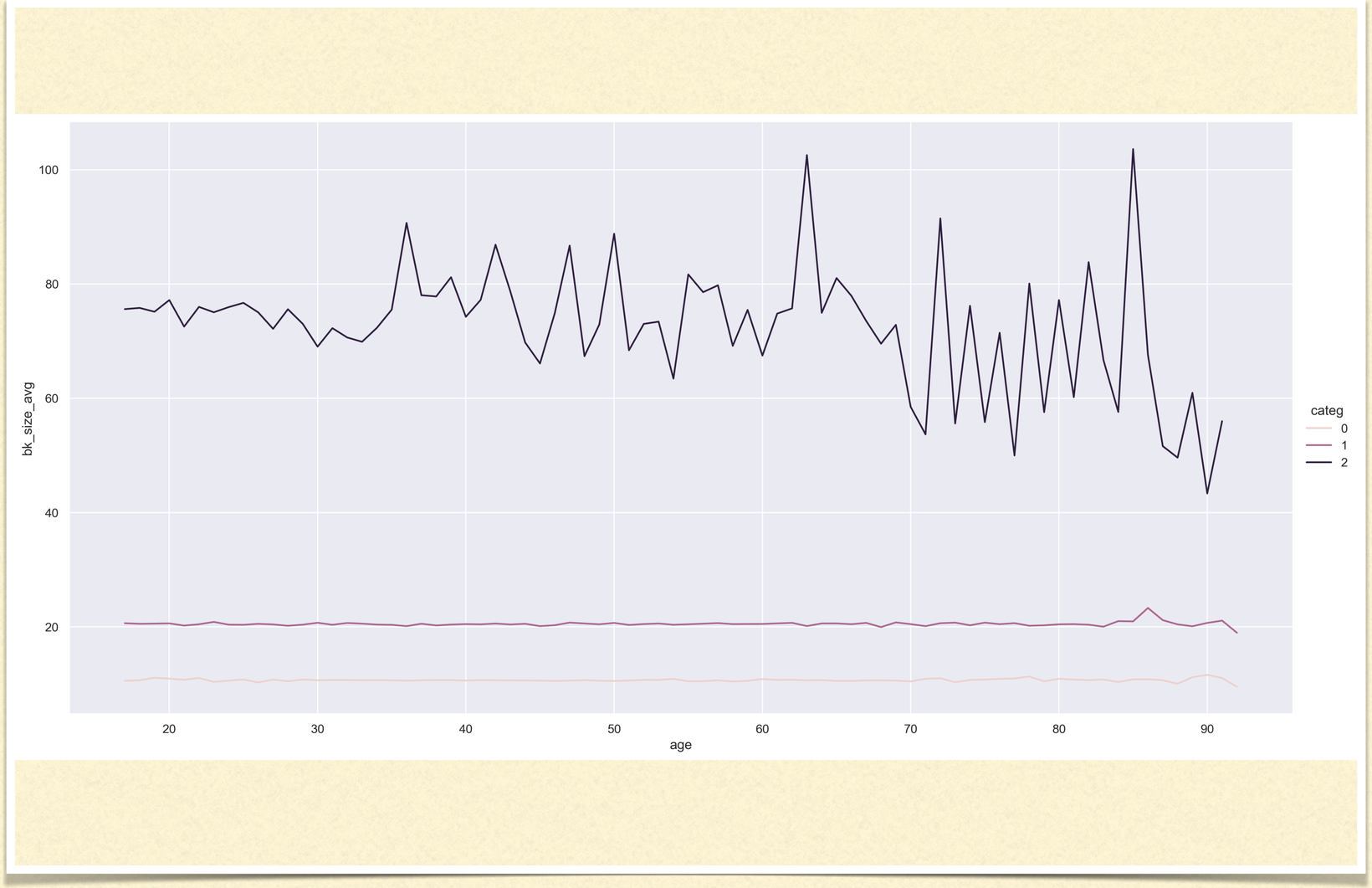


THE PURCHASE FREQUENCY (THE NUMBER OF PURCHASES PER MONTH FOR EXAMPLE)?

Considering age group with a difference of 20 years like 20S, 40S and so on.

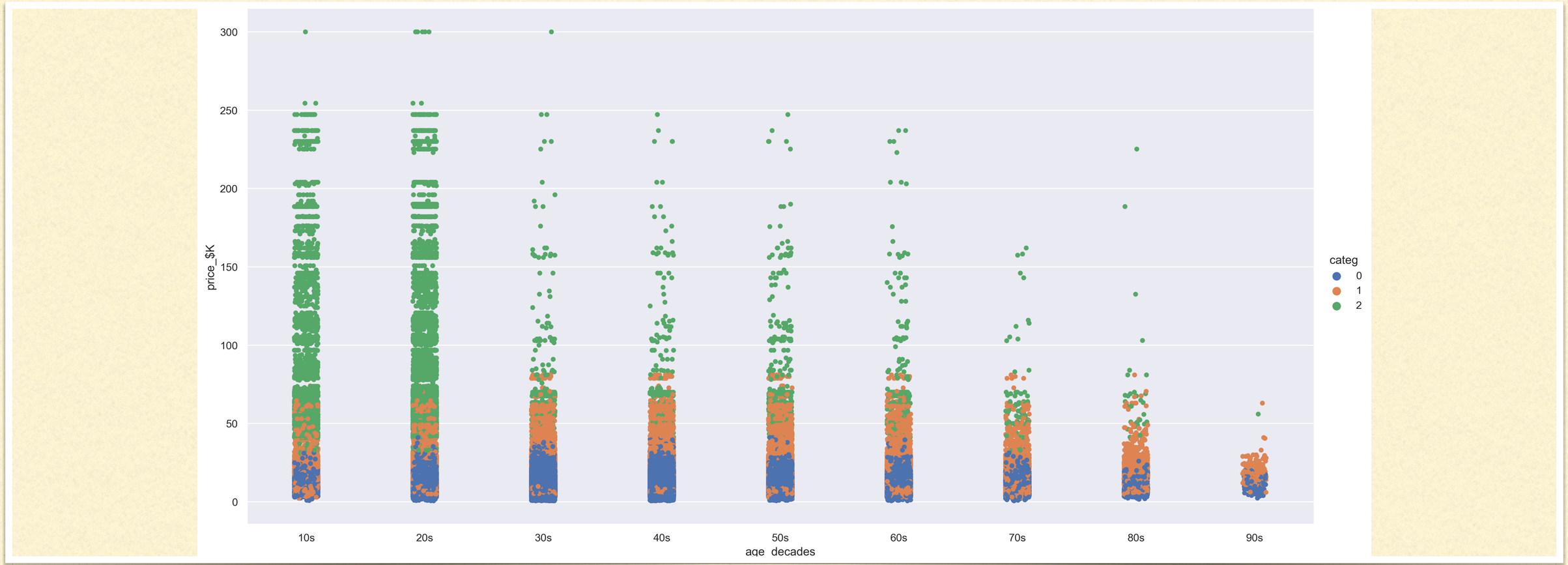
We see that during the whole year olds ones from 60s to 80s purchase far more products than young ones from 20s - 40s.

THE AVERAGE BASKET SIZE (IN NUMBER OF ITEMS)?



The category of product 1 and 2 have a close number of items when consider the basket size, both slightly increase at old age.

Category of type 0 has definitely more items with age range from 20 to 60.



CATEGORIES OF PURCHASED PRODUCTS?

If we consider the amount and how expensive the categories are related its value, and the age

we can see that less expensive category 0 and 1 is very dense aggregate and quite evenly distributed.