



---

Project Name: Customer Insight Project

Date: 15/03/2024

Analyst Name: Helen FitzGerald

---

**Contents:**

Population Flow

Consistency checks

Wrangling steps

Column derivations

Visualizations

Recommendations

**Citations:**

"The Instacart Online Grocery Shopping

Dataset 2017", Accessed from

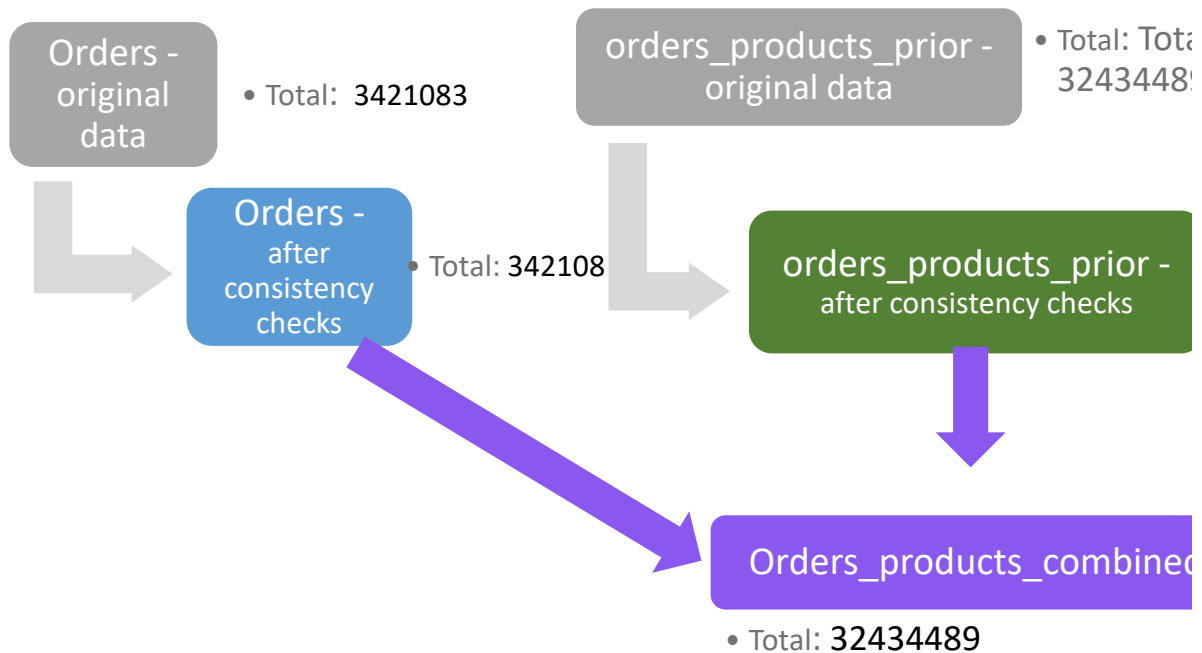
[www.instacart.com/datasets/grocery-shopping-2017](https://www.kaggle.com/datasets/instacart/instacart-online-grocery-shopping-2017)

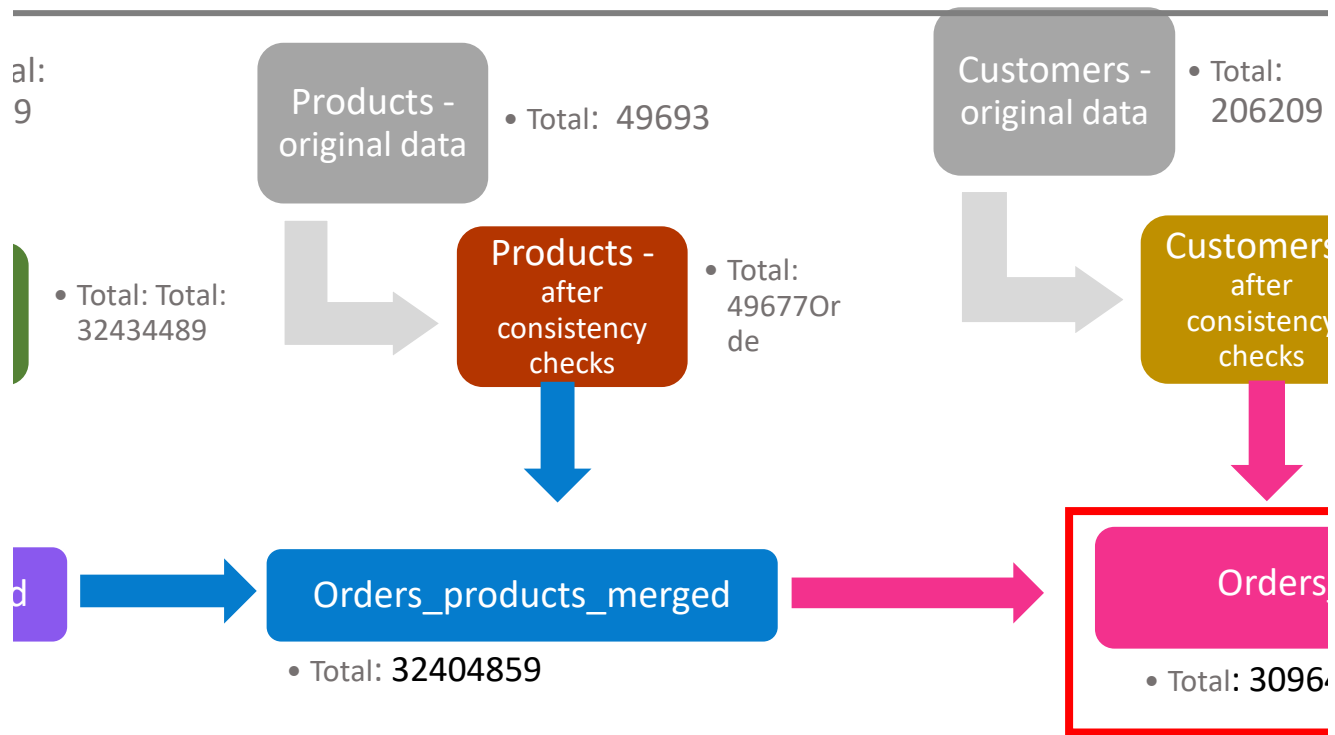
via Kaggle on 11/3/2024





## Population flow





---

S -  
y

- Total:  
206209

\_products\_all

4564



## Duplicates

None found.
-------------

5 duplicates were identified and all were dropped.

None found.

None found.

[illegible]



## Wrangling steps

[illegible]



Columns' type changed	Comment/Reason
	Unnecessary
	Security
	Security
er	Clarity
	Consistency re. naming conventions
	Consistency re. naming conventions
	Consistency re. naming conventions
	Consistency re. naming conventions
	Consistency re. naming conventions
	Consistency re. naming conventions
	Consistency re. naming conventions
Orders: user_id -> converted to str.	Consistency
Customers: first_name -> converted to str.	Consistency

Title page

## Column derivations and aggregations

---

Dataset	New column	Column/s it was derived from
ords_prods_merge	busiest_days	order_day_of_week
ords_prods_merge	busiest_period_of_day	order_hour_of_day
ords_prods_merge	price_label	prices
ords_prods_merge	freq_flag	median_days_between_orders
ords_prods_merge	spending_flag	mean_product_price
ords_prods_merge	loyalty_flag	max_order

ords_prods_all	region	state
ords_prods_all	exclusion_flag	max_order

Conditions
<p>Busiest: order day of week is 0 (Sunday) or 1 (Monday)</p> <p>Regularly busy: order day of week is 2 (Tuesday), 5 (Friday) or 6 (Saturday)</p> <p>Slowest Days: order day of week is 3 (Wednesday) or 4 (Thursday)</p>
<p>Most orders: 10 , 11, 14 , 15 , 13 or 12</p> <p>Average orders: 16, 9, 17, 8, 18 or 19</p> <p>Fewest orders: else</p>
<p>High-range product: &gt;15</p> <p>Mid-range product: &gt;5 and &lt;=15</p> <p>Low-range product: &lt;=5</p>
<p>Frequent customer: &lt;=10</p> <p>Regular customer: &lt;-20</p> <p>Non-frequent customer: &gt;20</p>
<p>High spender: &gt;=10</p> <p>Low spender: &lt;10</p>
<p>Loyal customer: max_order &gt; 40</p> <p>Regular customer: max_order &gt;10 and &lt;= 40</p> <p>New customer: max_order &lt;= 10</p>

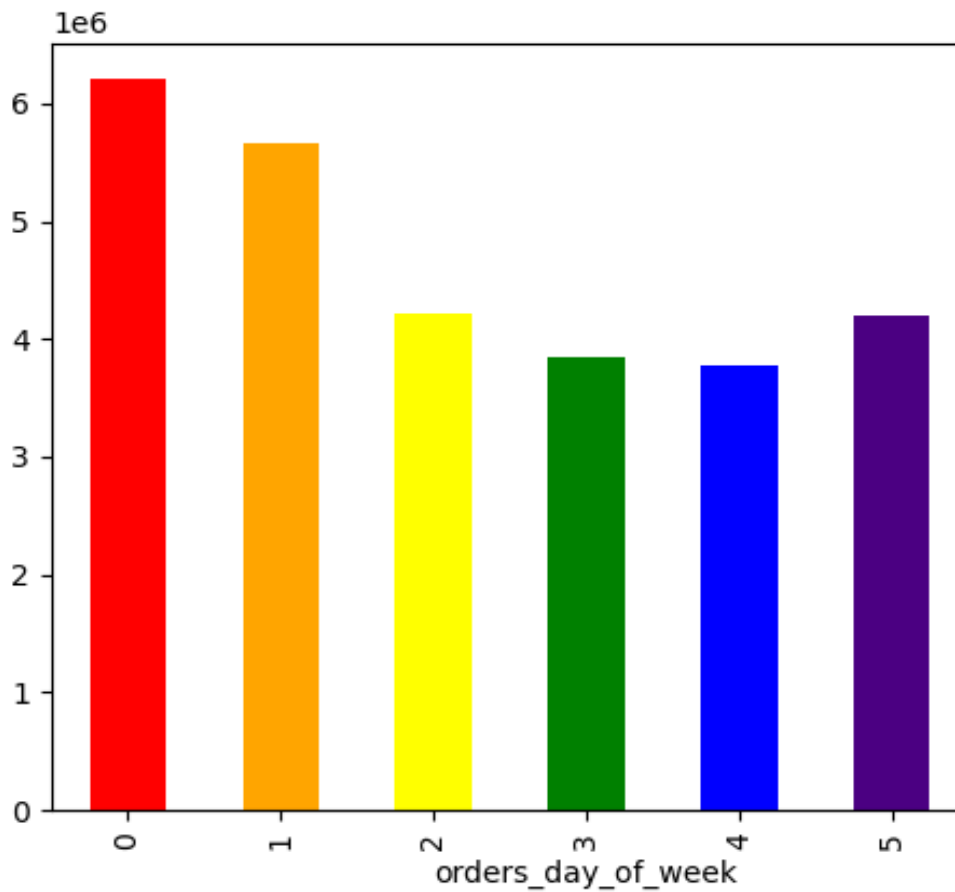
```
nw_list= ['Maine', 'New Hampshire',  
'Vermont', 'Massachusetts', 'Rhode Island',  
'Connecticut', 'New York', 'Pennsylvania',  
'New Jersey']  
mid_list= ['Wisconsin', 'Michigan', 'Illinois',  
'Indiana', 'Ohio', 'North Dakota', 'South  
Dakota', 'Nebraska', 'Kansas', 'Minnesota',  
'Iowa', 'Missouri']  
s_list = ['Delaware', 'Maryland', 'District of  
Columbia', 'Virginia', 'West Virginia', 'North  
Carolina', 'South Carolina', 'Georgia',  
'Florida', 'Kentucky', 'Tennessee',  
'Mississippi', 'Alabama', 'Oklahoma', 'Texas',  
'Arkansas', 'Louisiana']  
w_list = ['Idaho', 'Montana', 'Wyoming',  
'Nevada', 'Utah', 'Colorado', 'Arizona', 'New  
Mexico', 'Alaska', 'Washington', 'Oregon',  
'California', 'Hawaii']"
```

if max\_order is < 5, then exclusion flag = low  
activity

if max order is >=5, then exclusion flag =  
high-activity

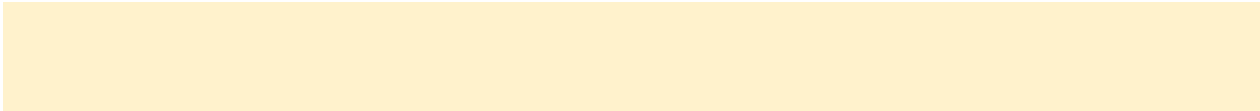

Title page

**What's are the busiest days of the week?**

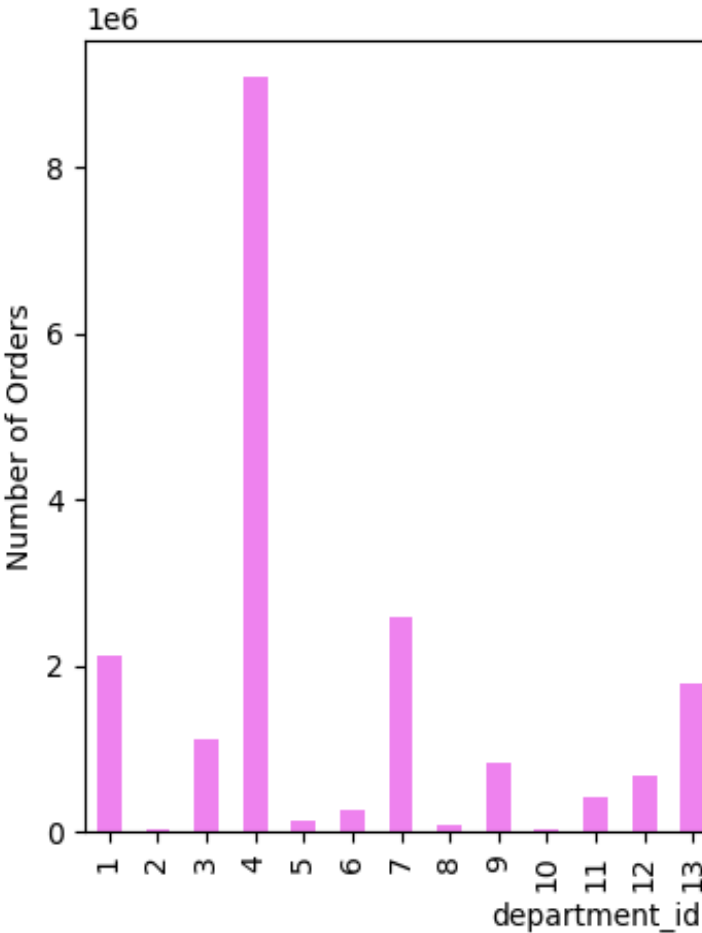


Sunday and Monday are the busiest days of th



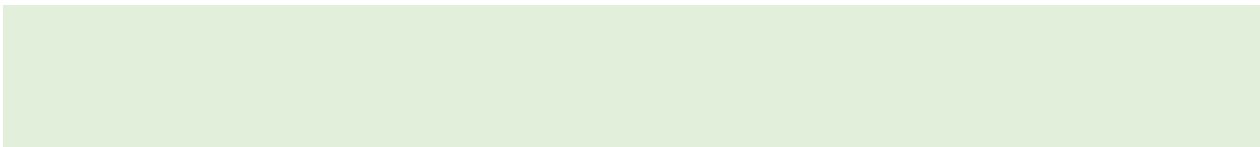


What departments perform

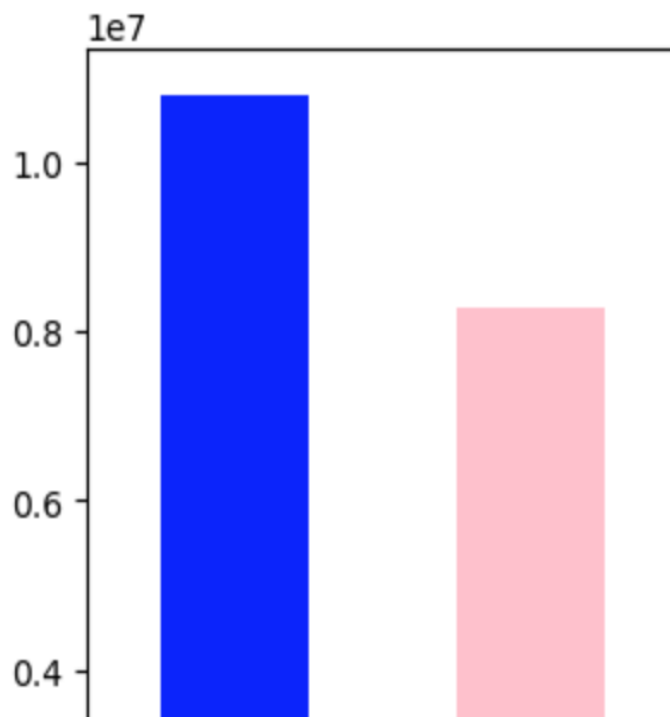


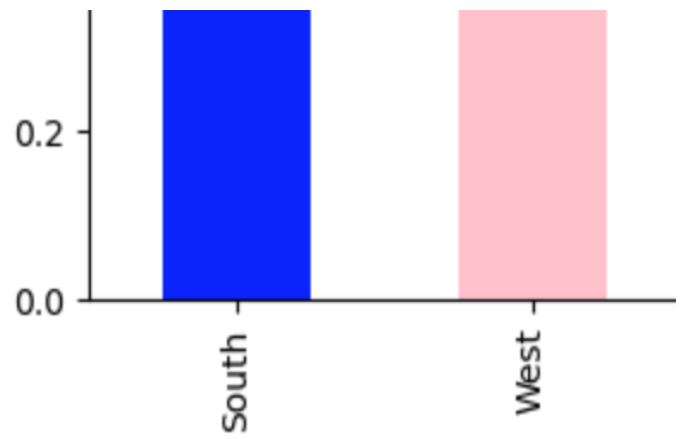
The 'Produce' department is the r  
'Dairy/Egg

The 'Other' category and the 'Bulk' c  
poorly.



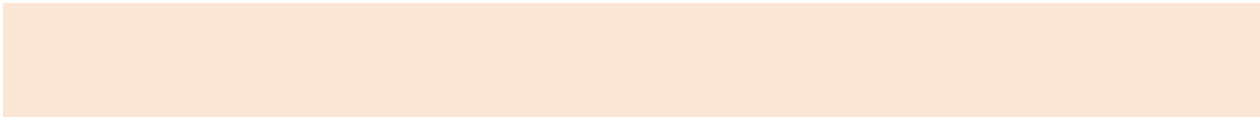
## How are customers



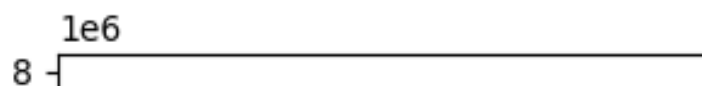


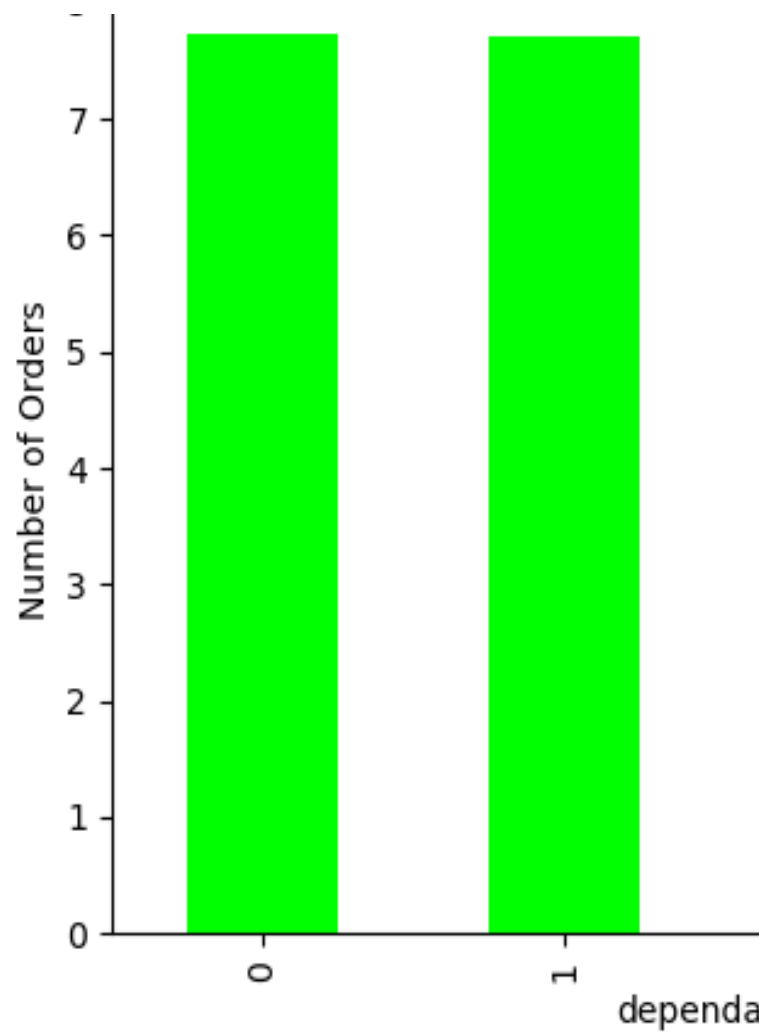
reg

Most customers are based in the Sout  
based in the No



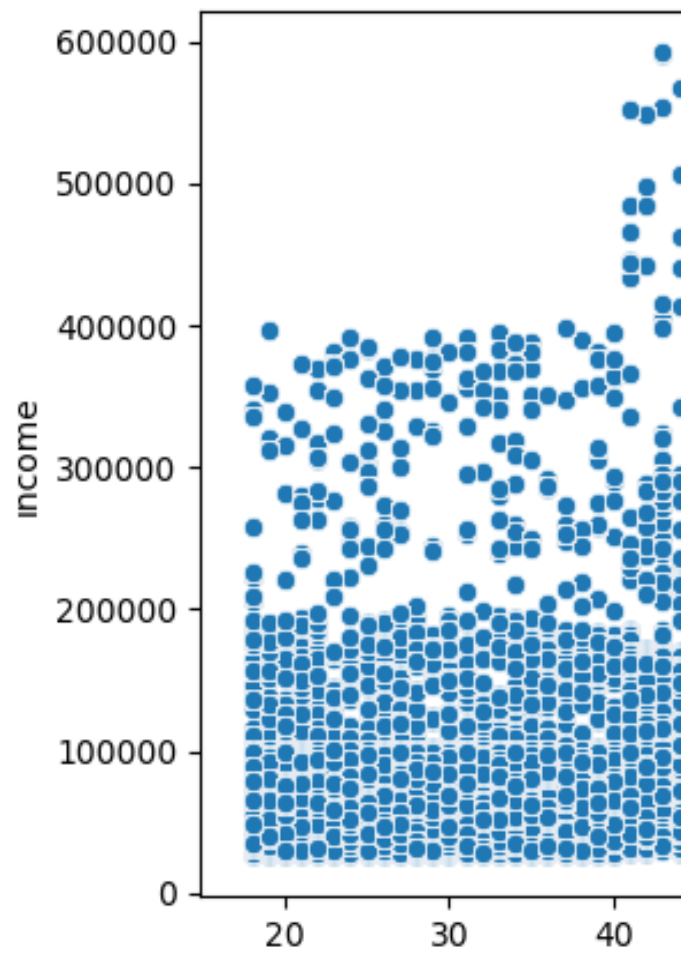
Does dependant count impact or





The quantity of dependants does not affect the customer's amount of orders.

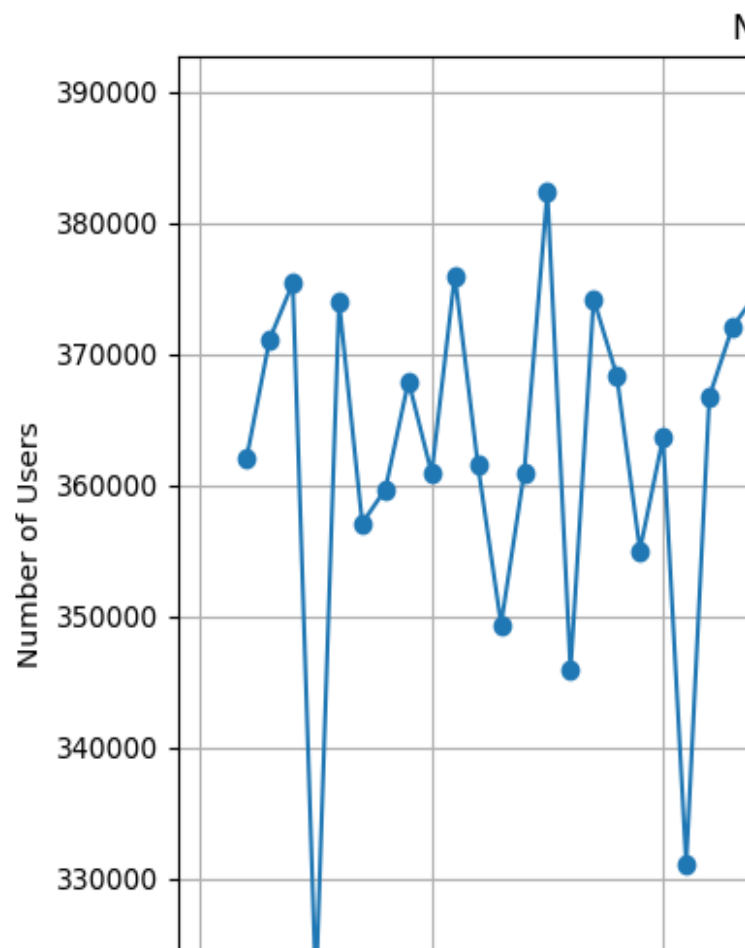
What's the relationship between

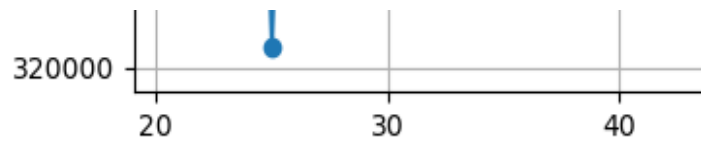


Answer: The highest dens

beneath 200,000 in income  
in the age range of 40-80,  
200,000-300,000. The in  
spread across all age ra  
finally, the income band o  
those aged betw

**So the majority, of customers are mai**



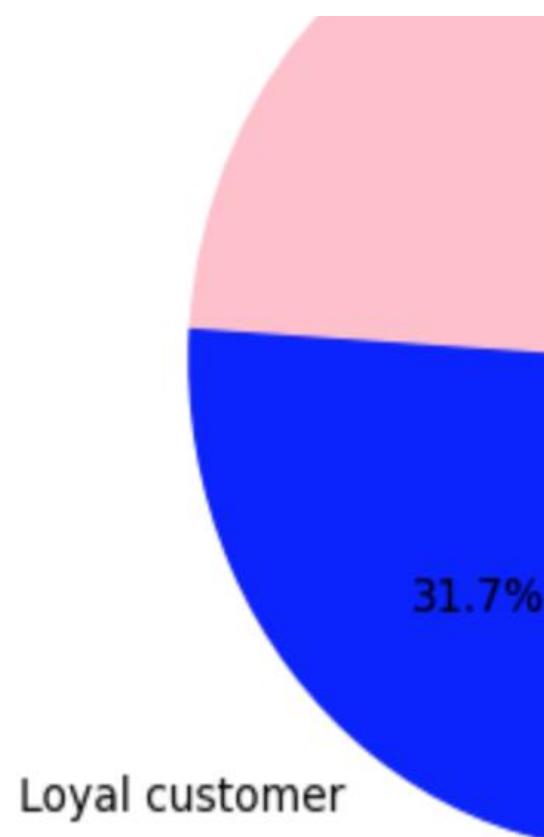


Many married custo  
typically be considere  
however, a signifi  
associated wi

**What is the current br**

Distribution



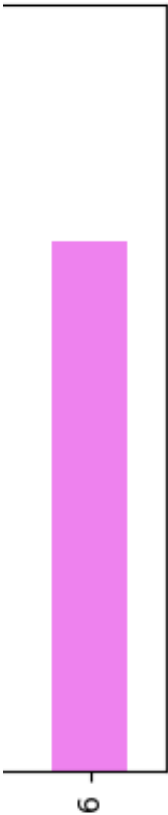


Our 'Regular' customer is promising to see so r  
vital that we incentivis  
(

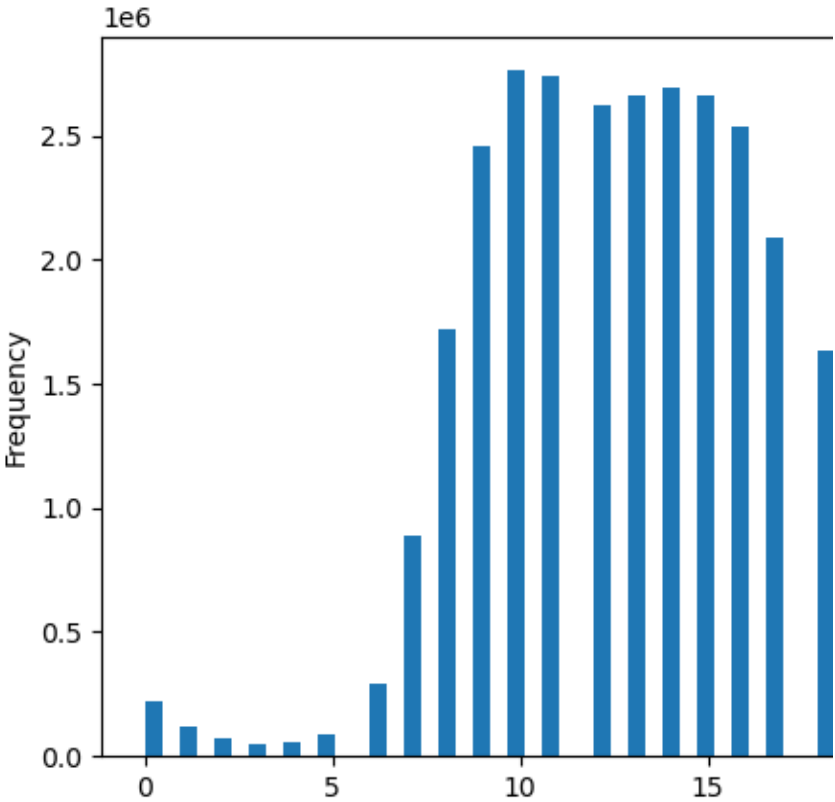


# Customer Habits

What's the busiest period of the c



e week.

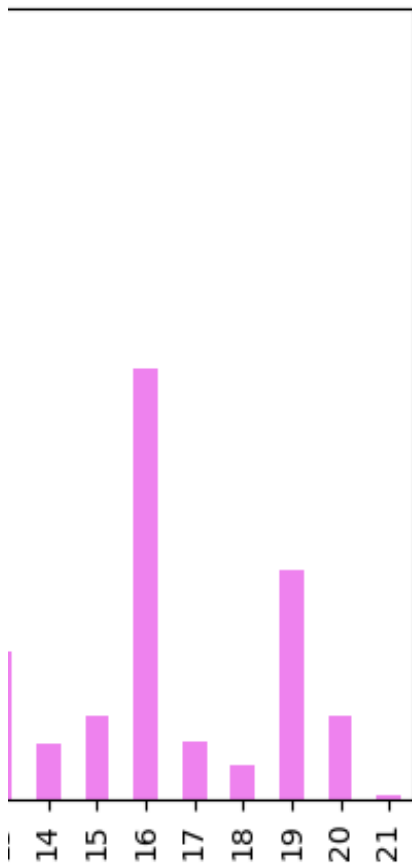


Between 9 am and 4 pm is the bus  
period.

Between 1 am and 5 am is the quie

# Product & Department B

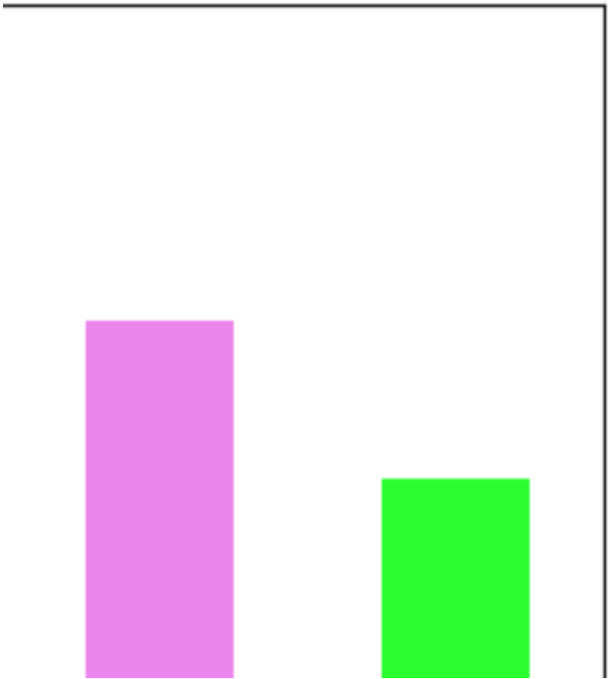
n well/poorly?

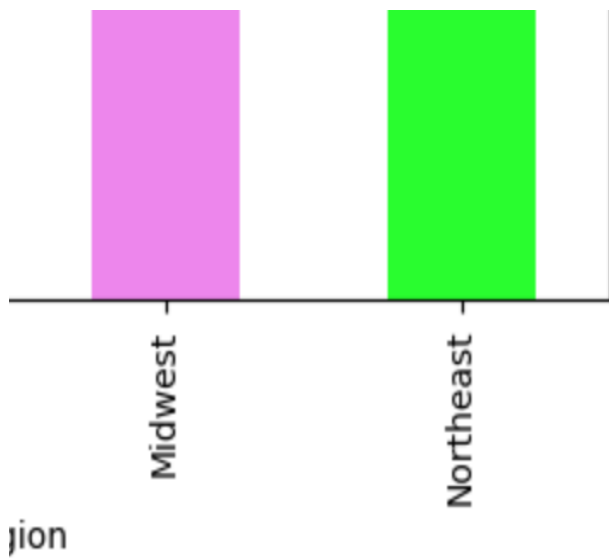


most popular, followed by  
s'.  
category perform particularly

# Regional Breakdo

divided by region?



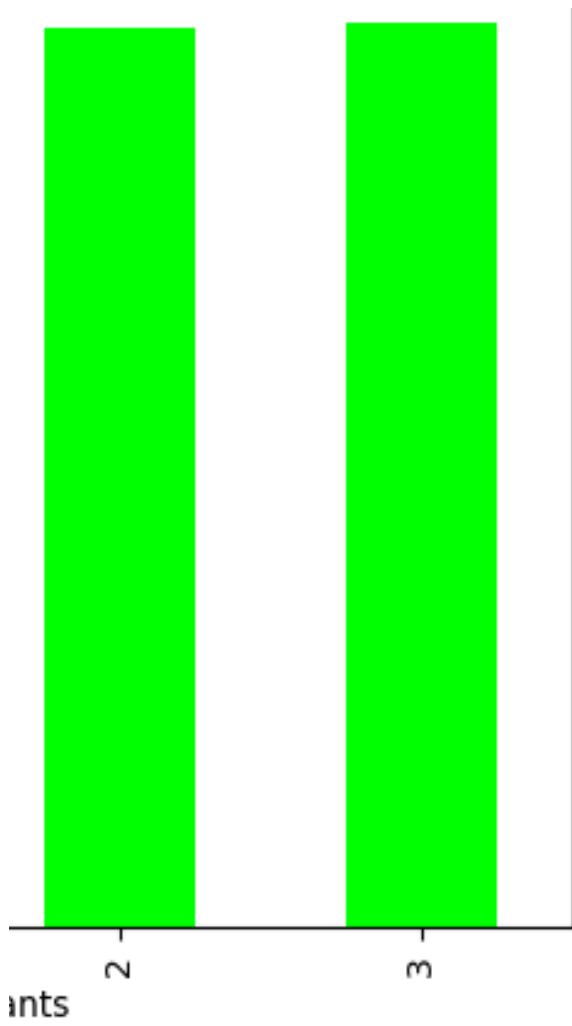


ch region, while the fewest are  
ortheast.

## Customer Breakdown

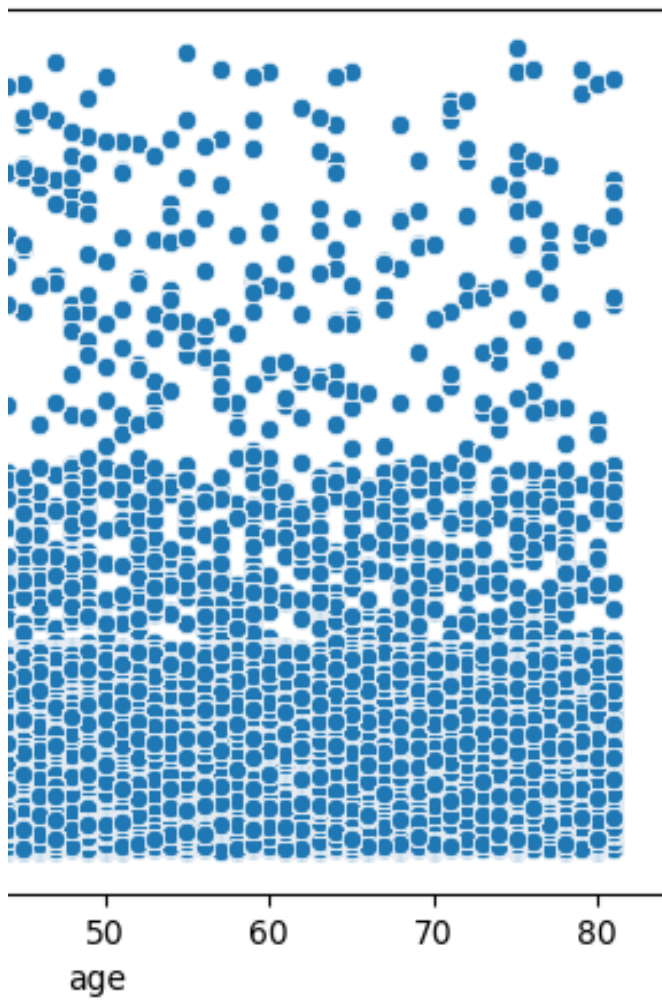
der count?





not appear to impact a given  
orders with Instacart.

How does customer age and income?

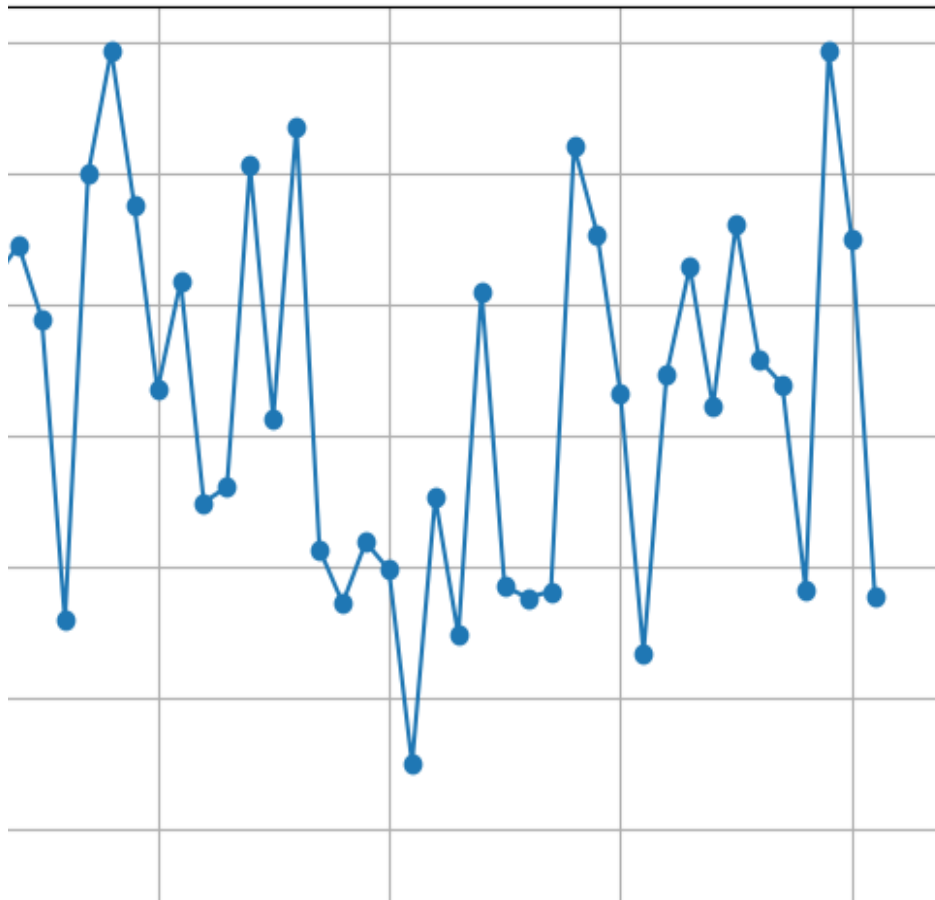


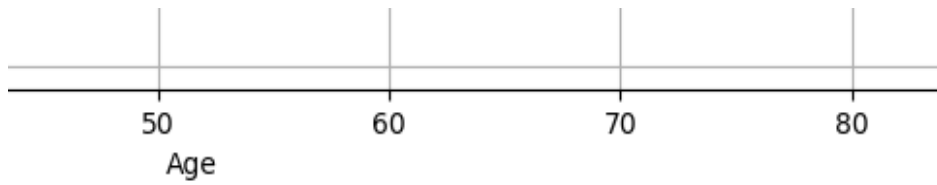
density of users (across age ranges) sit

re. The next clear clustering occurs for whom their income is between come band of 300,000-400,000 is ages, though less densely. And of 400,000-600,000 is reserved for reen 40-80 in this sample.

**ried. What ages are these married people?**

Number of Users by Age





omers are beyond the scope of what would  
d the standard age to parent young children,  
icant portion are within the ages usually  
ith having young children i.e. 26-35.

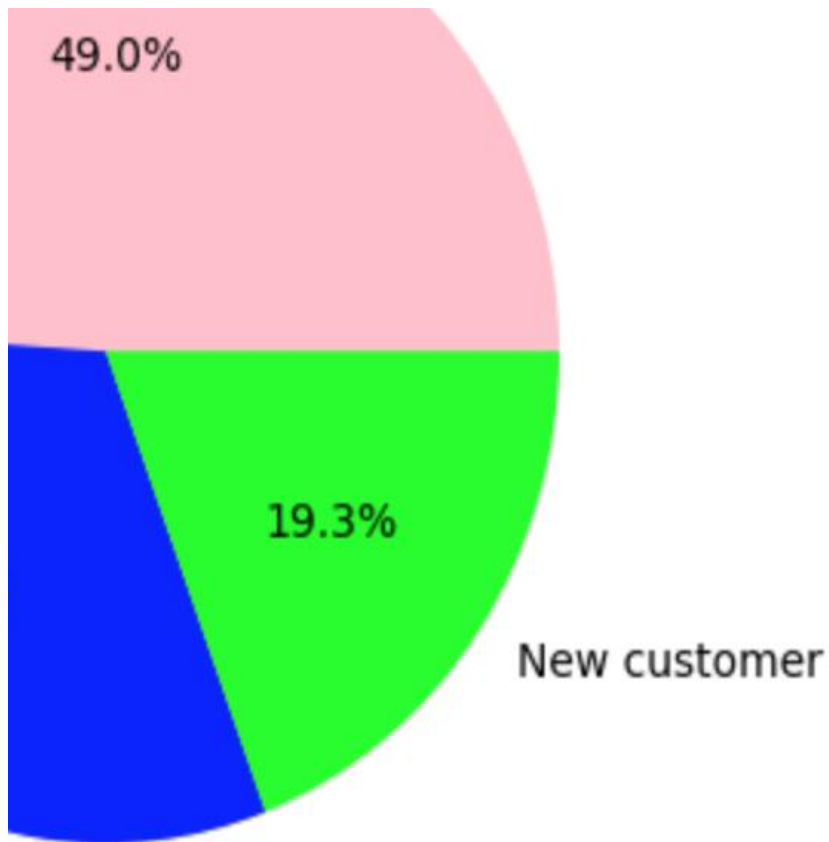
**breakdown of loyalty status?**

**of Customer Loyalty**

Regular customer





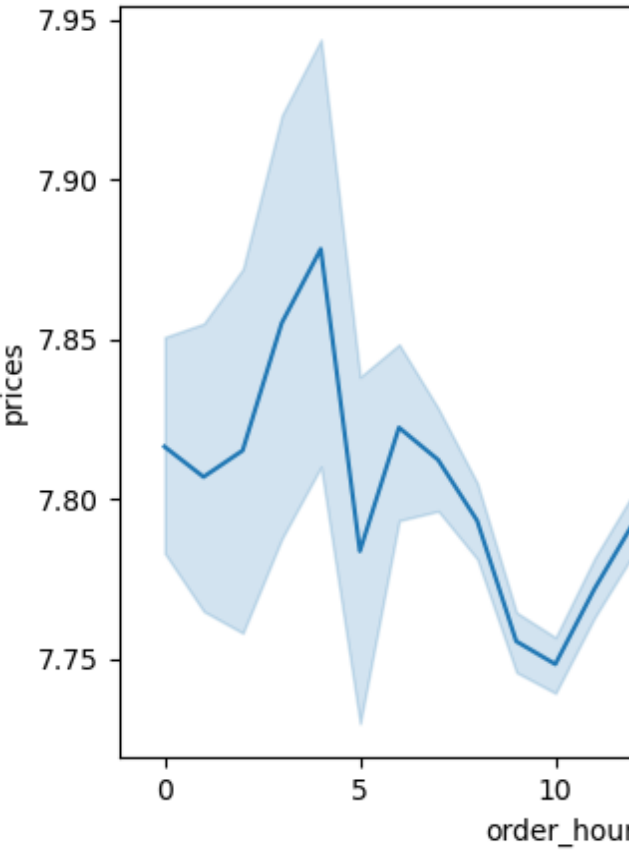
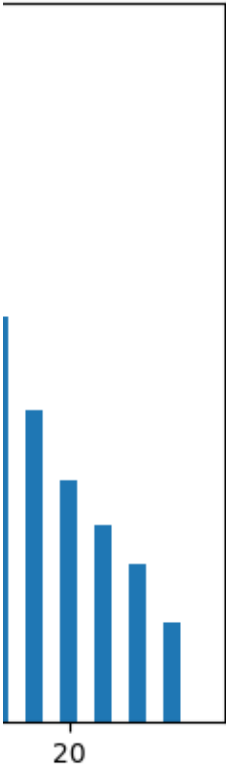


er count is quite healthy and it  
many 'New' customers, but it is  
se these new customers not to  
drop off.

# S

day?

When do people spend the



iest  
test

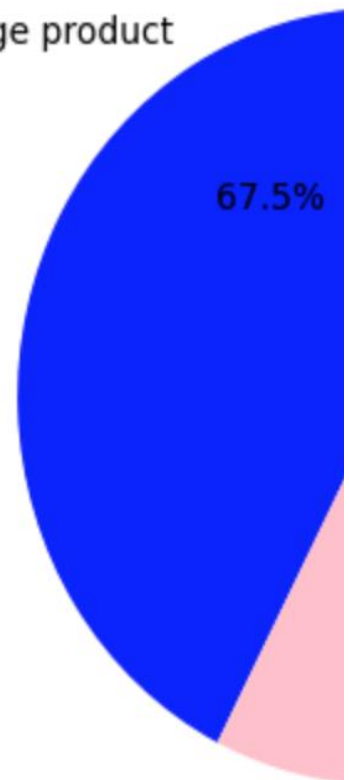
The highest priced items are so

# breakdowns

What is the price-range breakdown?

Mid-range product

67.5%

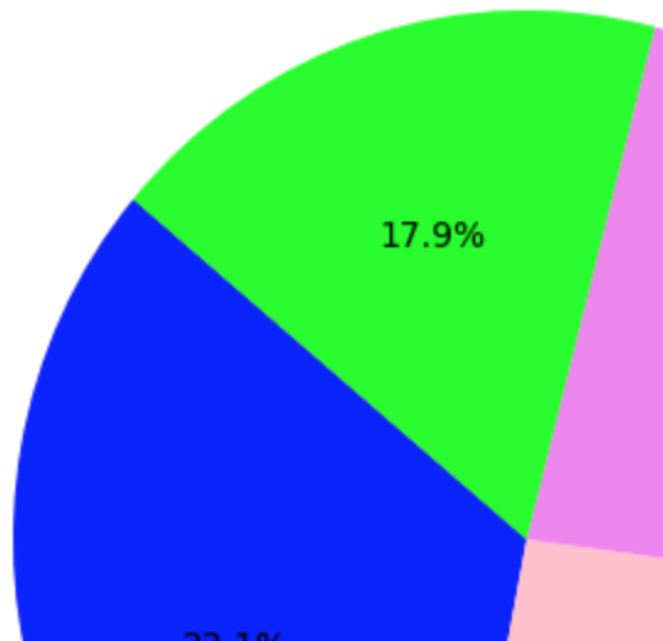


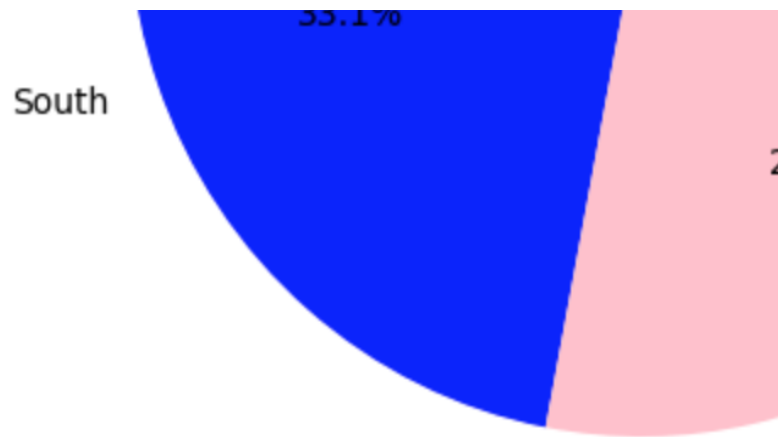
67.5% of stock is priced at the 'mid-range'  
only 1.3% of stock being

owns

Where are our loyal cus

Loyal Customers Breakdown  
Northeast



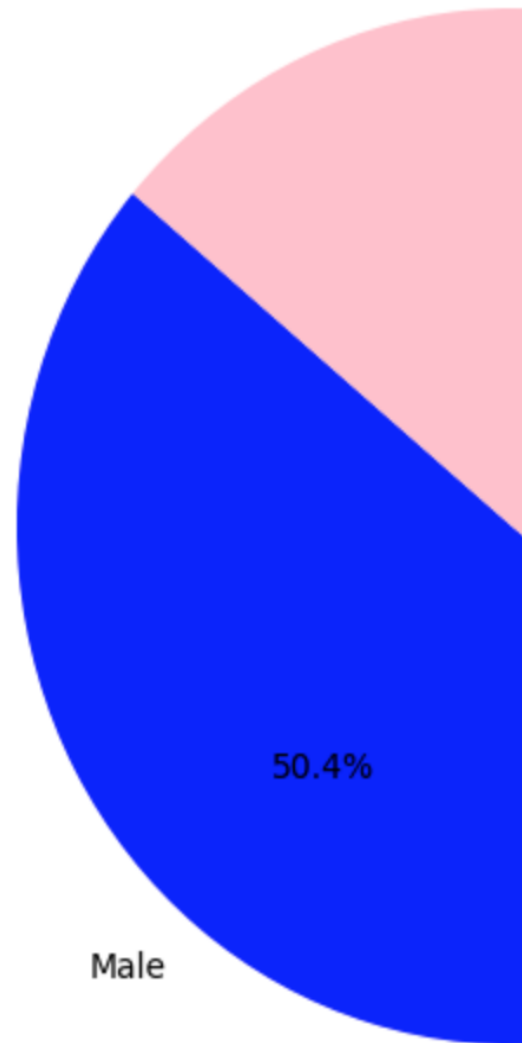


As well as hosting the most customers, the South is also the most loyal customers. This indicates that the South is more proportionate to customer loyalty.

## owns

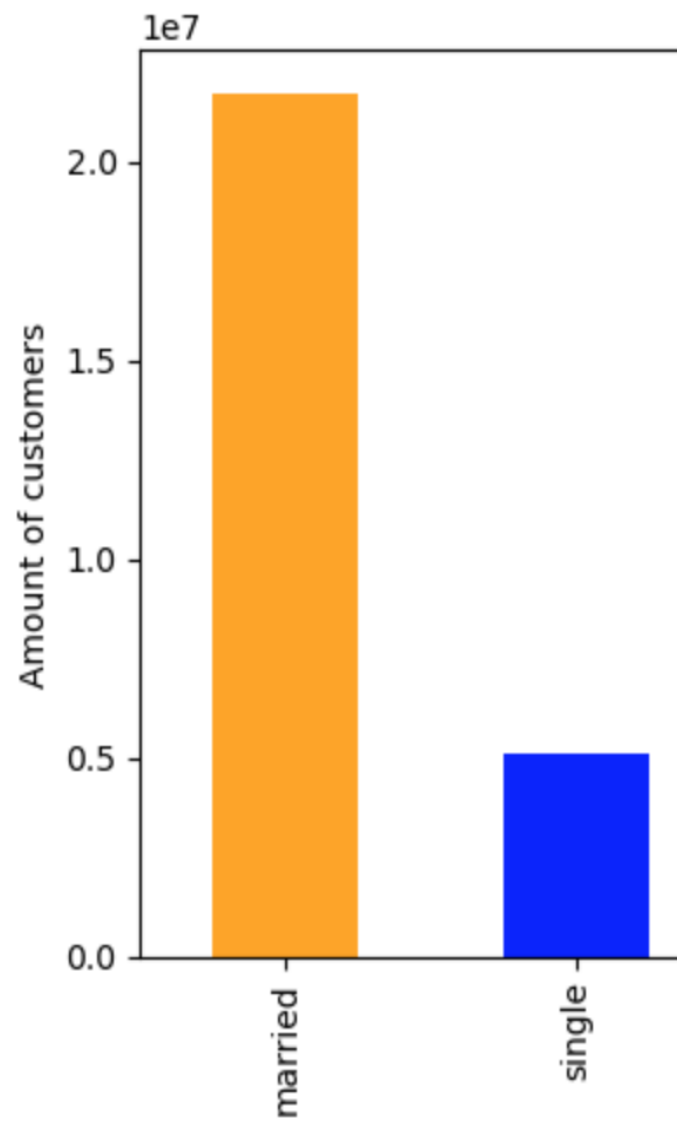
What is the gender breakdown of the customers?

Gender Breakdown of the Customers



The gender breakdown of o  
evenly split r

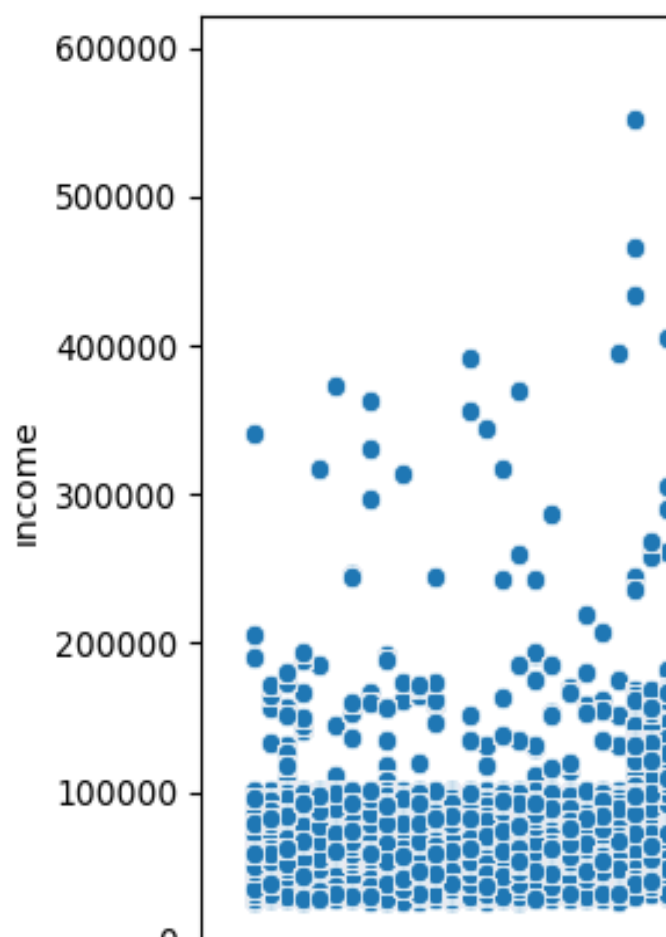
What is the family stat



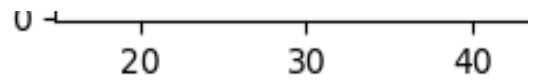
fam

The overwhelming maj  
mar

**What age and income leve**







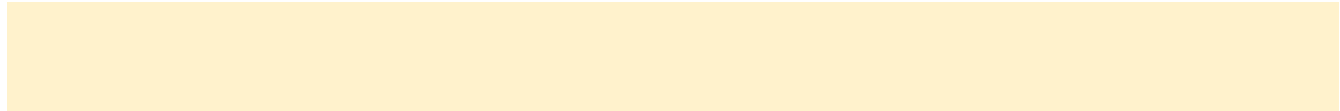
Interestingly, in comparison to the prior cus  
customers under 40 who earn up to \$400,C  
reach 'loyal' status are



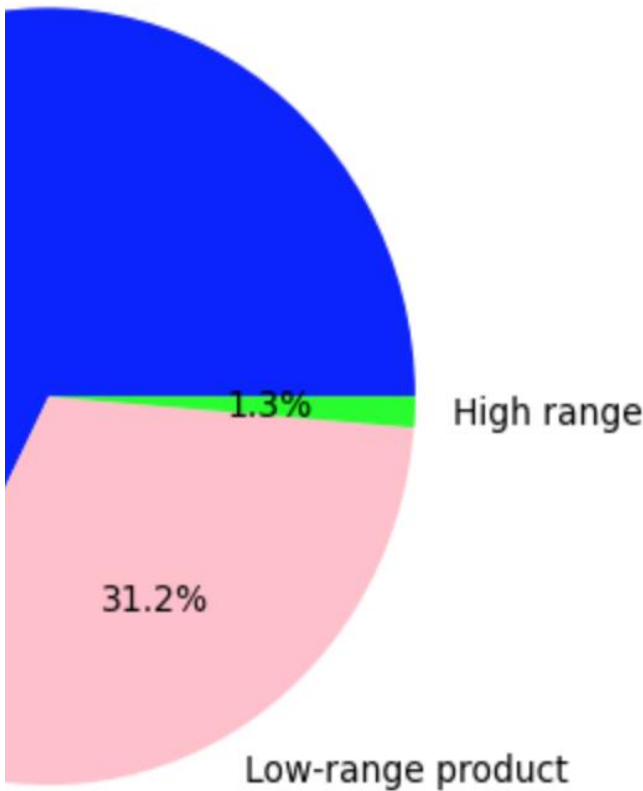
**e most amount of money?**



Id between 2 and 6 am



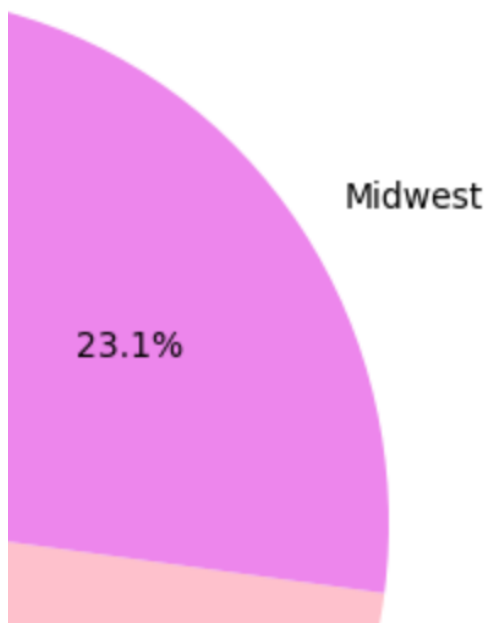
**breakdown of our products?**

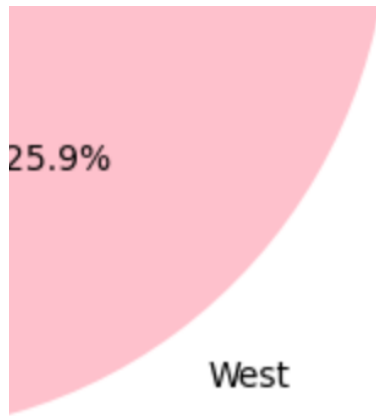


ange' category (>\$5 and <=\$15) with  
ng 'high-range' (>\$15).

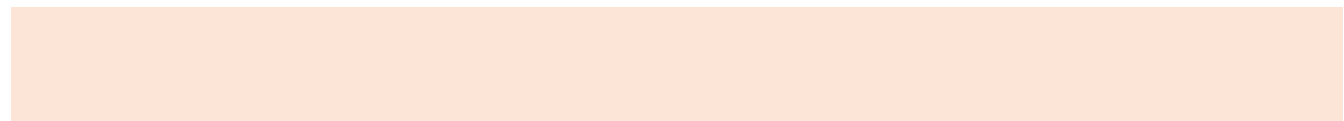
customers based?

wn by Region



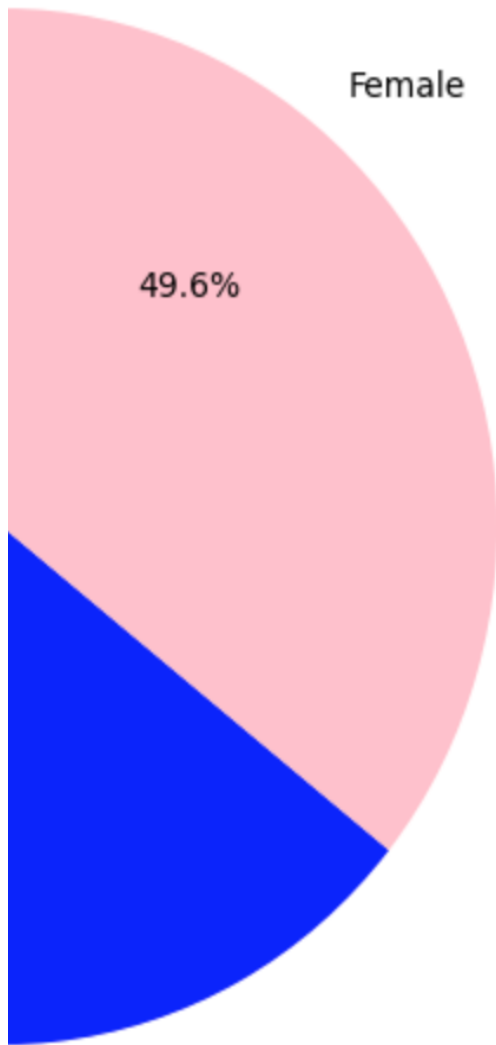


The South is also home to the most customer loyalty per state is per-base per state.

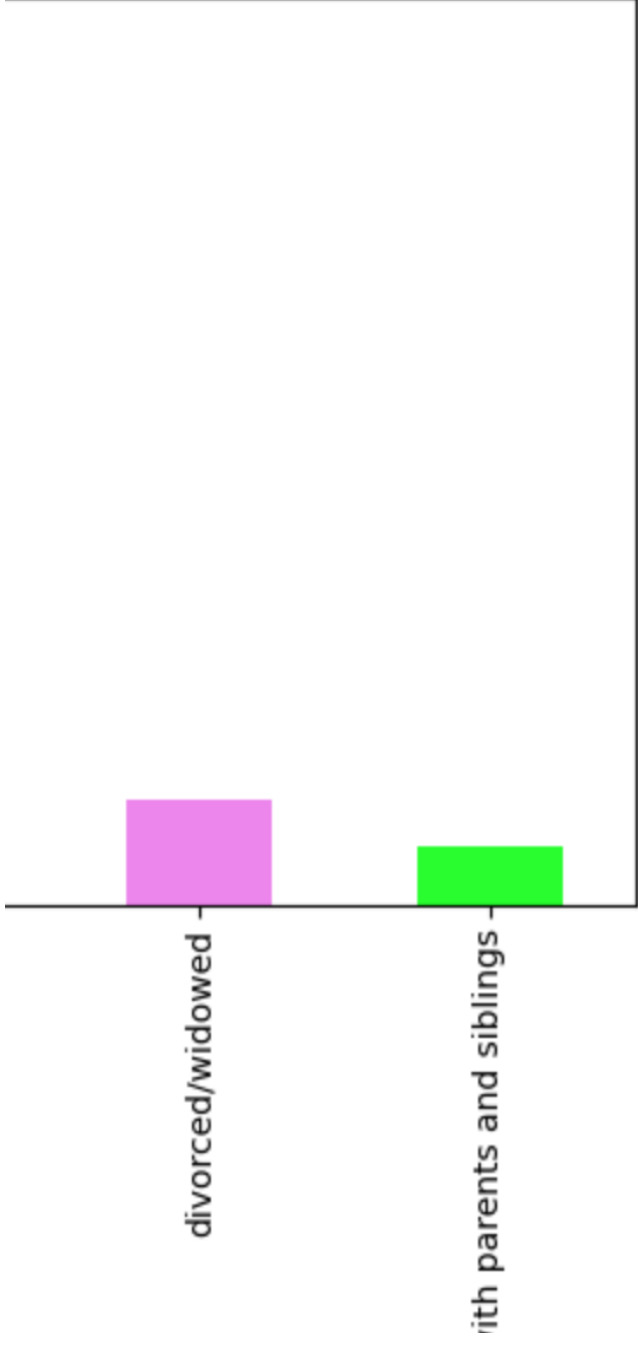


n of loyal customers?

of Loyal Customers



Our loyal customers is rather male/female.



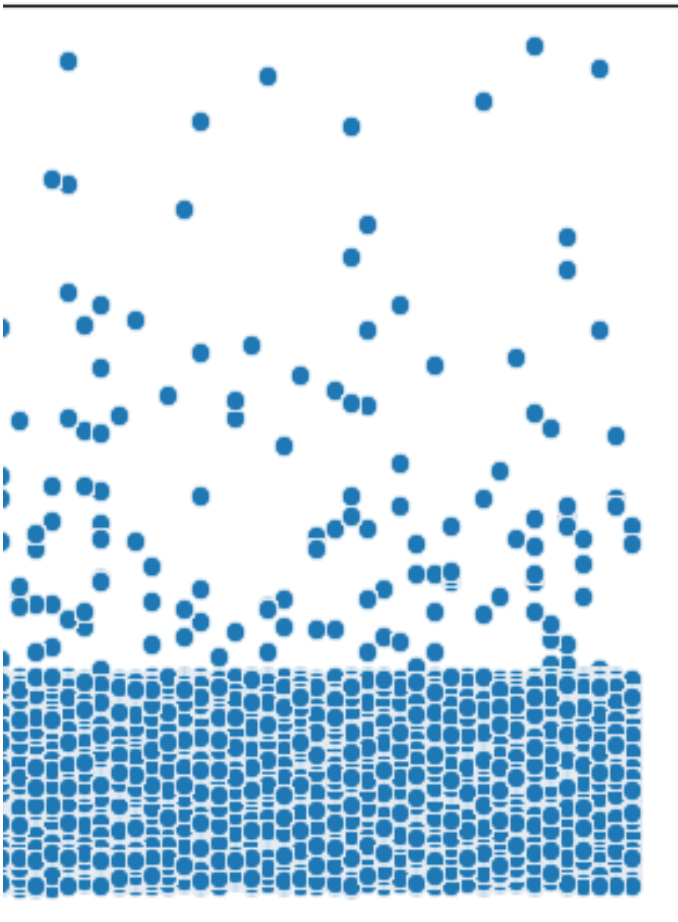


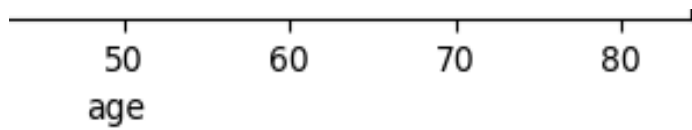
ily\_status

living w

majority of customers are  
ried.

What do our loyal customers have?





customer age to income chart, we see that despite having 100,000, the majority of customers in this age bracket who earn \$100,000 or less.

































## Recommendations

---

Additionally, I would advise to take advantage of

**2. Department-focus:** The 'Produce' and 'Dairy/Egg' departments c  
buying versatile s  
Furthermore, it is evident that 'dependants' are not a major influen

It may also be worthwhile reviewing what we categorise as 'C  
**3. Review inventory:** Product pricing has been divided into 3 categ  
we currently

**4. Regional adjustments:** It has been found that regional differenc

**5. Incentives:** There is a large amount of 'New' customers (19.3%  
customers. One way to achiev

Question	
1	<i>The sales team needs to know what the busiest days of the week and hours of the day are (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.</i>
2	<i>They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times</i>
3	<i>Instacart has a lot of products with different price tags. Marketing and sales want to use simpler price range groupings to help direct their efforts</i>

4	<i>Are there certain types of products that are more popular than others? The marketing and sales teams want to know which departments have the highest frequency of product orders.</i>
5	<i>The marketing and sales teams are particularly interested in the different types of customers in their system and how their ordering behaviors differ</i>

---

.

the qui

dominat  
staples s  
ce in cu

Other', a  
ories: lo  
do not h  
es do nc

5). It is n  
e this is

---




---

days.

et hours (1 am - 5 am) with more niche, targetted advertising e.g. herbal tea:

ie Instacart's sales, whereas 'Bulk' purchases remain low. Considering that th  
uch as produce/dairy/eggs should be emphasised. If such options do not cur  
stomer spending and the 'Babies' department has rather low sales. Therefore

(especially due to the high volume of married customers of child-  
s this may be a catchall term for miscellaneous items but it could be skewing  
w, mid, and high range. The overwhelming majority of products sit in the mic  
have an attractive selection for those seeking more luxurious options (e.g. hig  
ot significantly impact customer behaviour, for example, 'Produce' was the to

marketing stance without overly catering to specific geographi  
ecessary to incentivise these new customers to recognise the value of Instac  
through a loyalty system in which points are accumulated based on activity,

---

### Answer

Sunday and Monday are the busiest days of the week with Wednesday and Thursday being the quietest.

9am to 4pm is the busiest period of day and 3 am is the most quiet.

The highest priced items are sold between 2 and 6 am.

The following price ranges have been introduced:

High-range product: >\$15

Mid-range product: >\$5 and <=\$15

Low-range product: <=\$5

The 'Produce' department is the most popular, followed by 'Dairy/Eggs'.

Key takeaways:

- Most customers are married -> likely living in multi-person homes.
- Most customers are based in the South region, this is also where approx. a third of loyal customers live.
- Dependents (i.e. children) are not a major factor in account activity.
- The customer base is mostly evenly split between male/female.
- The lowest performing departments are 'Other' and 'Bulk'.
- 'Babies' and 'Pets' sales are underperforming drastically. Considering our prominent user base of millennials, it is unlikely that we have so few customers who would fall into these buyer categories. Therefore it's likely that we are not offering them good deals on items purchased frequently (such as diapers or dog food).

---

s etc. to help insomniacs relax or coffee to boost early risers/those who

the majority of customers are married and likely sharing the home with children. If these products currently exist for customers, it should be explored.

Therefore, I would recommend investing more into market-research relating to customer needs and bearing age.

Based on the survey results. We should evaluate if products within 'Other' could be more targeted to a mid-range tier. Perhaps we should consider broadening our offerings with higher quality cuts of meat, wine, or electronics).

Top performing department regardless of location. Due to this, we can identify key demographic factors.

Part in order to return to the platform, as well as to express gratitude to customers, resulting in freebies, discounts or other promotional benefits.





orking late.

others, the benefits of bulk-

o the demographic of parents

accurately catalogued.

hin the high-range bracket as

embrace and employ a wide

oward 'Regular' and 'Loyal'