instacart

[Instacart Grocery Ba

Project Name: Instacart Grocery Basket Analysis

Date: 4/20/24

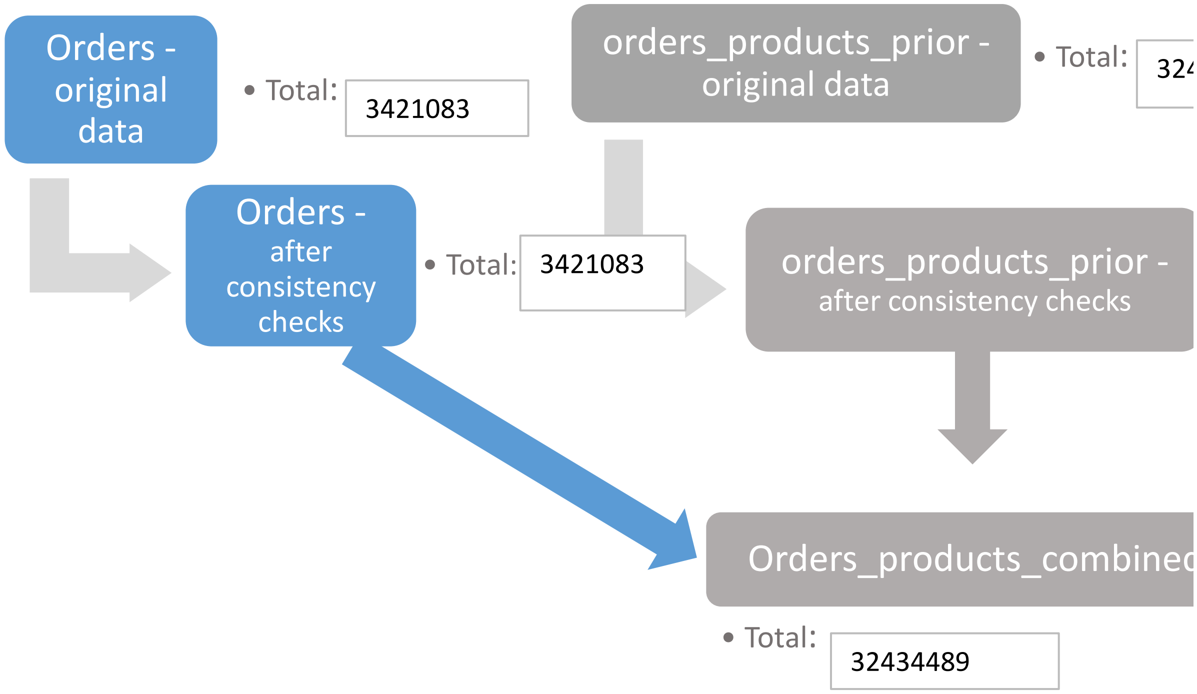
Analyst Name: Nancy Kolaski

Contents:

"The Instacart Online Grocery Shopping Dataset 2017", Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle on [4/20/24].

sket Analysis]

Population flow



Orders - original data

• Total: 3421083

orders_products_prior - original data

• Total: 324

Orders - after consistency checks

• Total: 3421083

orders_products_prior - after consistency checks

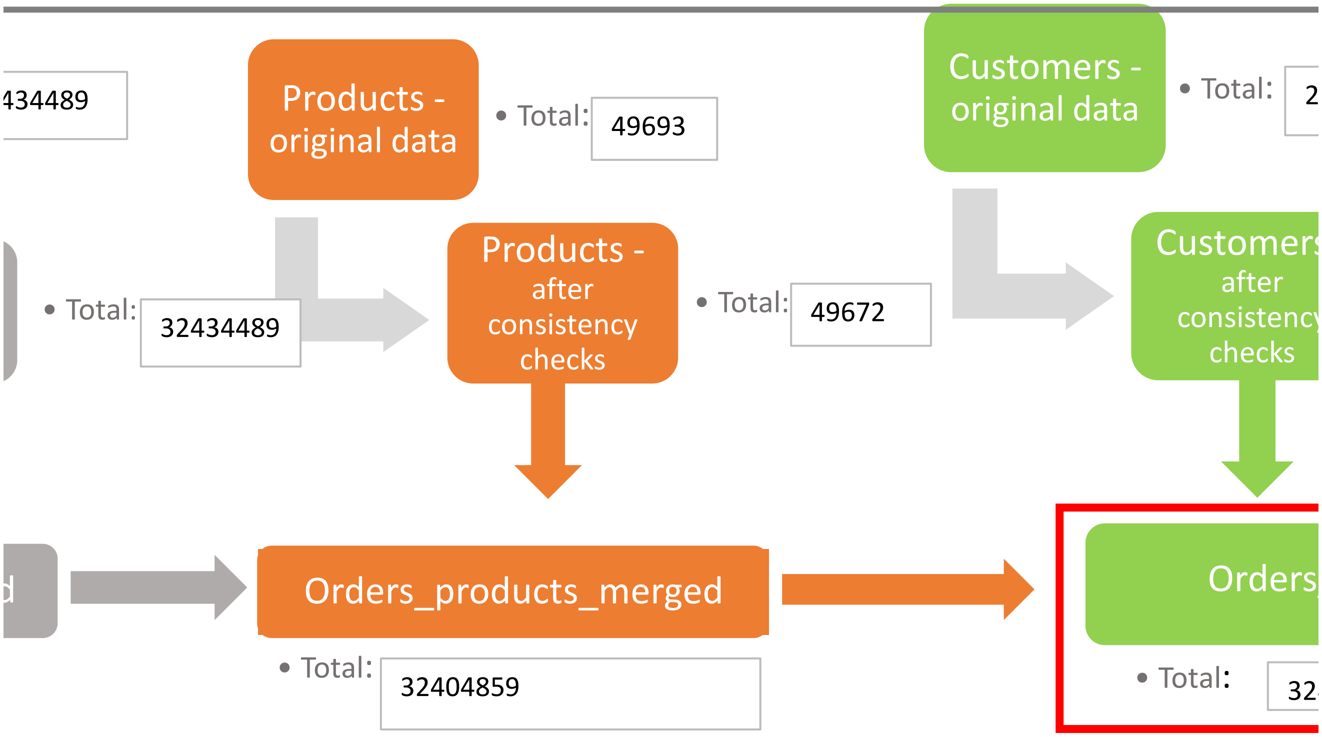
Orders_products_combined

• Total: 32434489

1.) The grey boxes in the first row of the population flow represent the origin imported the data set into Jupyter.

2.) The second row of boxes (coloured) represents the data sets **after** you ma conducting these operations. This offers a visual oveview of how the data *flo*

3.) The third row, where also the arrows are coloured, represents the merge that you end up with the final dataset (in the red box). Keep in mind the final



434489

Products - original data

- Total: 49693

Customers - original data

- Total: 2

- Total: 32434489

Products - after consistency checks

- Total: 49672

Customers after consistency checks

d

Orders_products_merged

- Total: 32404859

Orders

- Total: 32

ıal data sets as they were when you downloaded them. In the Total fields you need to a

anipulated them, e.g., removed missing values and duplicates. In the Total fields you ne
ws throughout the data consistency checks.

s you performed between the datasets. In the Total fields you need to add the count of
l dataset should be without exclusions (based on the exclusion flag).

Title page

206209

s -

y

- Total: 206209

Exclusion flag

Condition: max_order < 5
Obervations to be removed: 1440295
Final total count of order_products_all: 30964

_products_all

404859

add the count of the rows when you

eed to add the count of the rows after

f the rows in the merged datasets, so

4564

instacart

Consistency checks

Dataset	Missing values	Missing values treatment
orders.csv	206209 missing entries for 'days_since_prior_order'	created a flag because this is too large to discount
products.csv	16 missing entries for 'product_name' (49677)	
orders_products_merge	turn outliers (all values over 100 dollars in the price column) into NaNs as this cleans data # and we want it saved in this export (not just in the visualizations folder)	
customers		

Title page

Duplicates
dropped 5 duplicates (49672)

instacart

Wrangling steps

Columns dropped	Columns renamed
eval_set from orders.csv	
	'order_dow' renamed to 'orders_days_of_week
	order_hour_of_day' to 'order_time'
'Unnamed: 0.1', 'Unnamed: 0'	
	Surnam' to 'surname', 'Gender' to 'gender', 'STATE' to 'state', 'Age' to 'age', 'fam_status' to 'marital_status', 'n_dependants' to 'dependants'
	instacart_mege_2.p

Columns' type changed
orders.csv
'
df_prods (Script 4
customers.csv (Script
first_name'
ords_prods (Lesson 4
user_id'
cust_ords_prods (Leso
order_id', 'product_id', 'department_id' to string
kl (Lesson 4.10, merged dep
department_id" as string

Comment/Reason
eval_set' was dropped as it is irrelevant to current data. It represents prior Instacart data only
column nams changed for clarity
.6)
Columns were dropped as they did not contain valuable information
4.9.1)
changed to lower case for consistency of use during wrangling process, adjusted spelling errors, and changed names to make more sense for my own clarity when looking at the data
changed 'first_name' to string because it contained mixed data types.
4.9.1)
'user_id' colum is typed as object/string in 'customers', but as int64 in 'ords_prods'. I will change the 'user_id column in 'ords_prods' to match that type as it makes more sense and we want to combine datasets
n 4.9.1)
changed to string as that makes more sense as these are identifying numbers, not to be calculated
partments into cust_ords_prods))
for consistencyof dataytpe, same as above (as departmnt_id' from departments dataset showed up as object

Title page

instacart

Column derivations and aggregations

Dataset	New column
orders_products_merge	price_label:(turned into price_range_loc)
	busiest_day
	busiest_days
	busiest_period_of_day
	loyalty_flag
	frequency_flag
instacart	region
instacart_highspender_dep_merge	age_flag
	living_flag
	shop_time
	mean_orders
	dep_popularity
instacart_merged_flagged	spending-flag

Column/s it was derived from	Conditions
price_range	**High range product** > 15 **Mid range product:** > 5 & <= 15 **Low range product (10126321):** <= 5
orders_day_of_the_week	**Busiest day is** 0 **Least busy is** 4 **Regularly busy :** All other values
orders_day_of_the_week	**Busiest days:** isin([0,1]) **Regularly busy (12916111)**: isin([2,5,6]) **Slowest days (7624336)**: isin([3,4])
order_hour_of_the_day	**Fewest orders** : isin([23, 6, 0, 1, 5, 2, 4, 3]) **Average orders** : isin([17, 8, 18, 19, 20, 7, 21, 22]) **Most orders** : isin([10, 11, 14, 15, 13, 12, 16, 9])
max_order(user_id & user_order_number	**Loyal customer :** max_order > 40 **Regular customer :** max_order <= 40 & > 10 **New customer :** max_order <=10
days_since_prior_order (median)	**Non-frequent customer** : order_frequency > 20 **Regular customer :** order_frequency <= 20 & > 10 **Frequent customer** : order_frquency <=10
state	**'state' within 'region' defined by United**
age	**Youg Adult** : >=18 & <=29 **Adult :** >=30 & <=44 **Middle-Age Adult** : >=45 & <= 59 **Senior (10574504)**: >= 60
dependants	**Alone:['dependants'] == 0, 'living_flag']= 'Alone'** **With Family (['dependants'] > 0, 'living_flag'] 'Alone'**
order_time	**Early Bird shops between 5-8am** **Night Owl shops between 8pm-5am** **Regular is the default for the other times**
order_number (mean)	the mean of each order_number
mean_orders	**Not Popular if mean_orders are <=18** **Regularly popular if mean_orders <=21** **Most Popular if mean_orders >=21.1**
average prices	**Low spender:** average_price < 10 **High Spender:** average_price >=10

Title page

Flags/label Frequencies

Variables.	**count**
price_range_loc--	
Mid_range product.	20891771
Low-range product	9674840
High-range product	397953
busiest_day--	
Regularly busy	21430960
Busiest day	5906610
Least busy	3624994
busiest_days--	
Regularly busy	12349739
Busiest days	11320296
Least busy days	7294529
busiest_period_of_the_day------------------------	
Most ordesr	20180856
Average orders	9550810
Fewest orders	1232898
loyalty_flag--	
Regular Customer	15876776
Loyal Customer	

Flags/label Frequencies

Variables.	**count**
region--	
South	10311139
West	7927227
Midwest	7261413
Northeast	5464685
age_flag--	
30s & 40s	9730686
Older Adult	8195544
Middle Age	7220731
Below_30	5817603
living_flag---	
With Family	23224883
Alone	7739681
shop_time_flag--	
Reular	24908263
Night Owl	3168547
Early Bird	2887754
dep_popuarity---	
Not popular	16165087
Regularly bought	14700298

Visualizations

NOTE: All of the relevant visualizations are included in the next tab
did not show any insights, and were therefore left out of the 'Rec

The sales team wants to know what the busiest days of the week and hours of the day are in order

The chart below shows that Saturday(0) and Sunday(1) are the busiest days of the week,

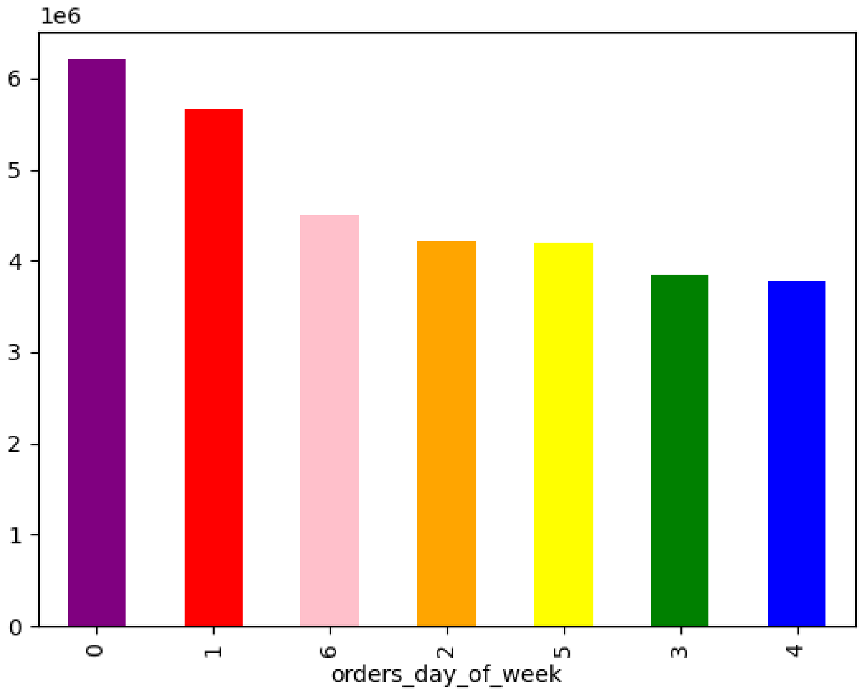
They also want to know whether there are particular times of day when people spend the most m

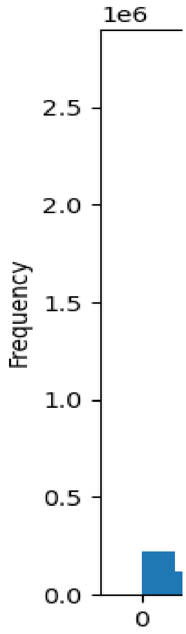
To it's right, the histogram shows the busiest hours of the day for Instacart sales is mid da

INSIG

Schedule more Ads for mid-week (Tuesdays & Wednesdays) since these are th

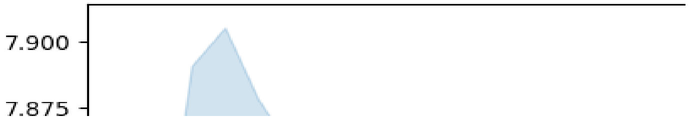
Seeing as busiest times for Instacart orders are ocurring mid-day at 12-3pm, it is beneficial to adve

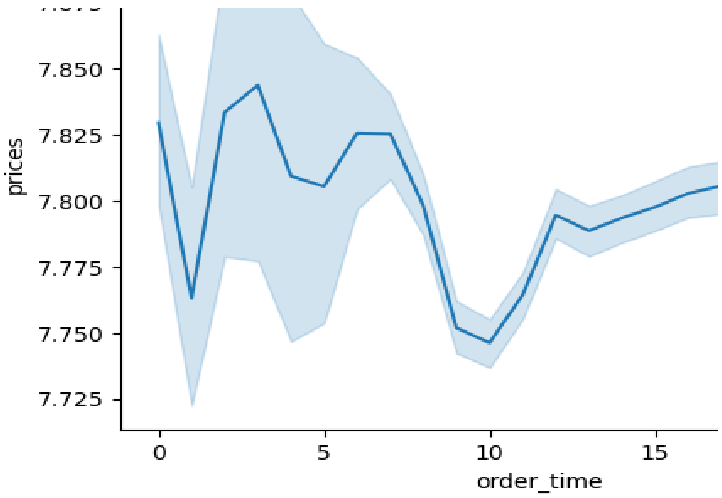




INSIGHT:

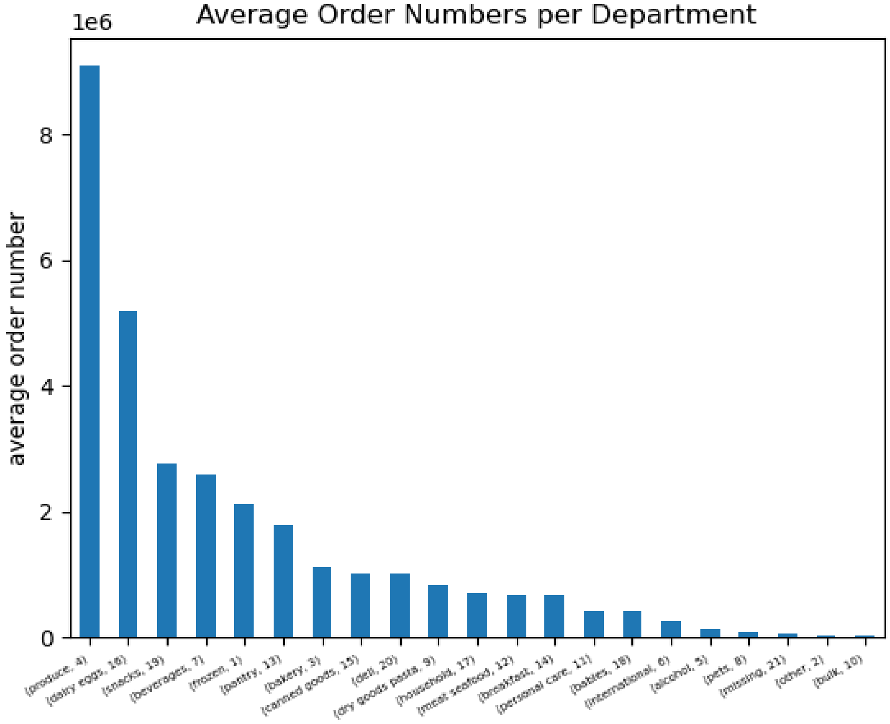
THis line chart below shows that most higher priced ite
early morning hours (around 2-3am).   Could be due to
shift stocking up before they go home, or those wh
deciding to prepare for the day. In either case, it is int



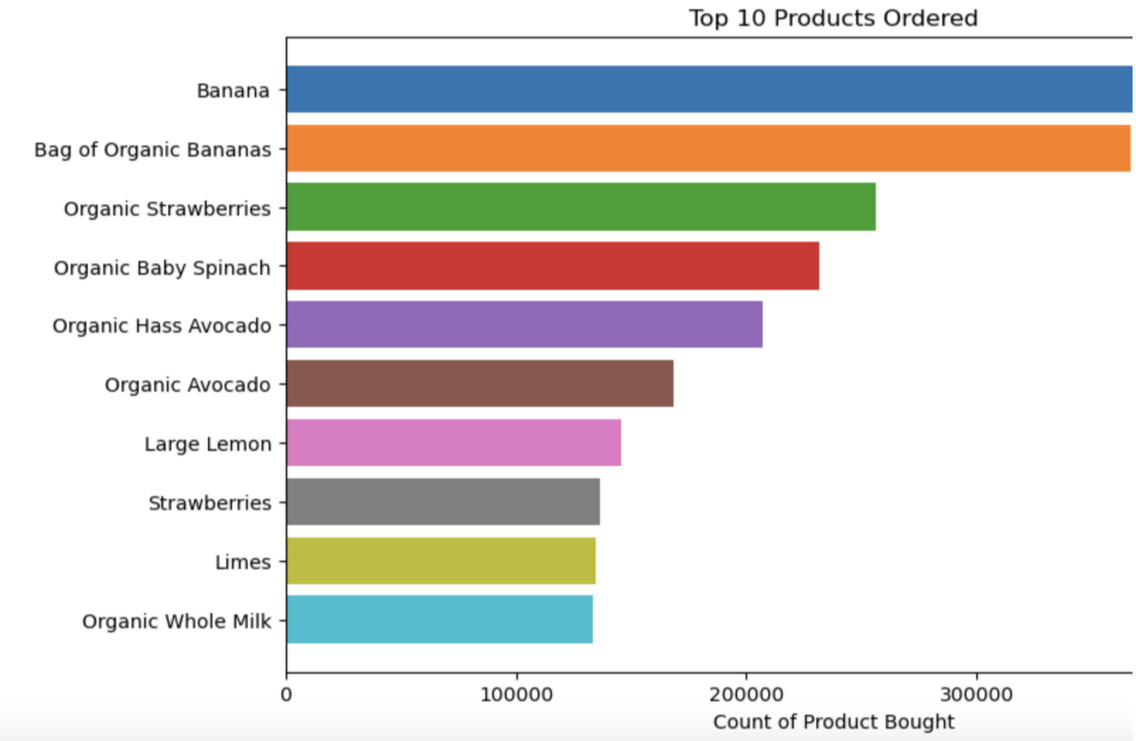


Instacart has a lot of products with different price tags. Marketing and sales want to use simpler p
Are there certain types of products that are more popular than others? The marketing and sales te

INSIGHT:
This bar chart below shows the information I want in a clear format, demonstrating the
for each department! **Produce** is the most popular by far with a mean of 9,079,273 av
followed by 5,177,182 orders for **dairy/eggs**. **Snacks**, **beverages**, **frozen**, and pantry c
following those top contenders as most popular departments. The least popular depa
international, alcohol, and pets.

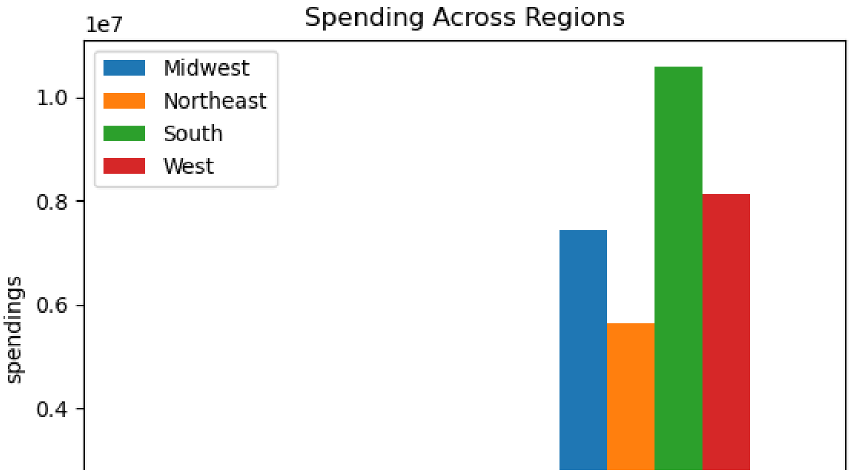


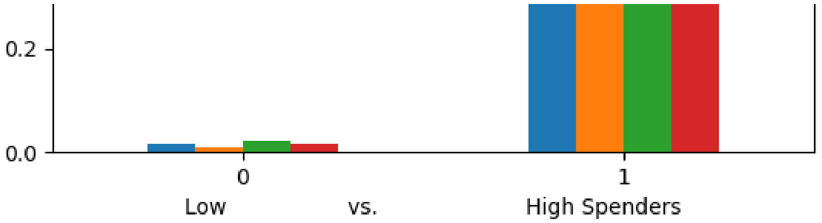
INSIGHT:
This bar chart below shows the top 10 products ordered: ranked with regular bananas as
banans, organic strawberries, organic baby spinach, organic Avocados, lemon, regula
organic whole milk. Intersting to note the popularity of organic products. Implication w



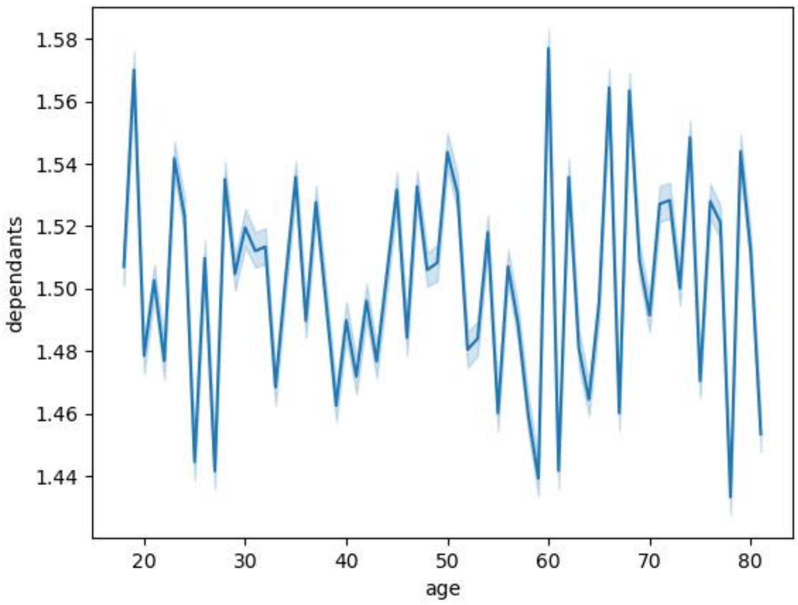
The marketing and sales teams are particularly interested in the different types of customers in th

INSIGHT:
The chart below shows that **Southern Regions** generate the most of Instacart's spendings. This is followed by Western Regions, then Midwest Regions, with Northeast spending the least. This pattern is the same for both groups of people



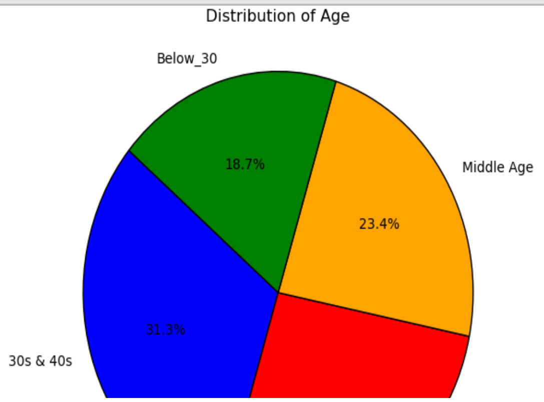


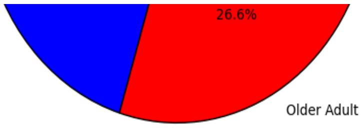
The line chart below shows that there is no definitive pattern here to suggest that there is any correlation between age and family situation (or number of dependants).



This scatterp
this scatter
$400,000 t
200,000 to 3

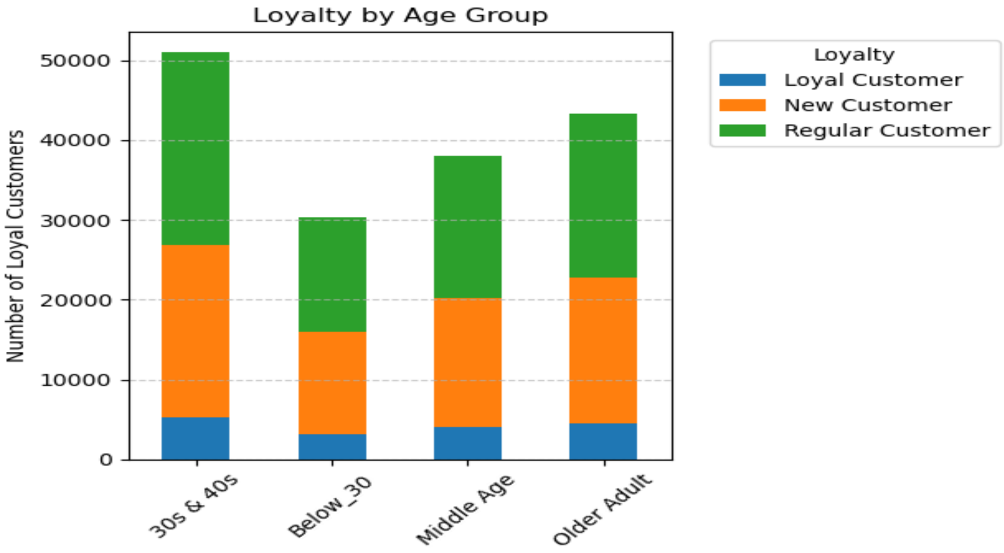
Majority of customers are in their 30s and 40s (shown by the blue), howev
the percentage gap is not a huge difference between other ages.



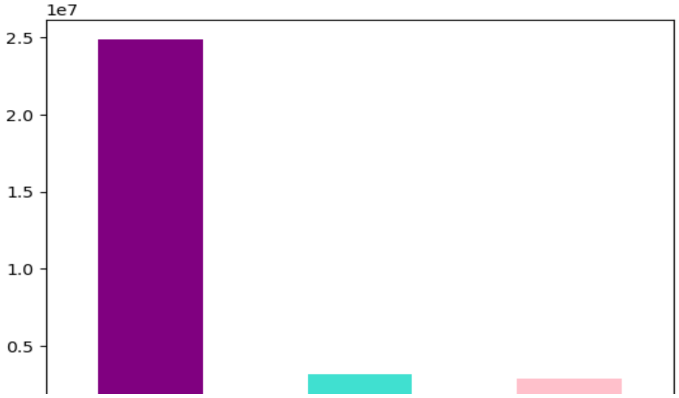


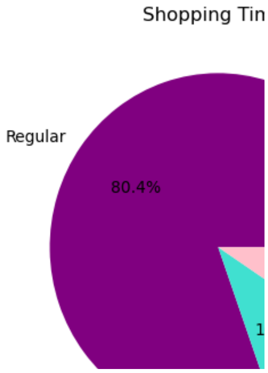
INSIGHT:
The **30s/40s** age group tends to rank the highest in all aspects of customers according to this bar chart: They make up 1)the most loyal group to Instacart, 2 the most regular customers, and 3) they make up the majority of new customer:

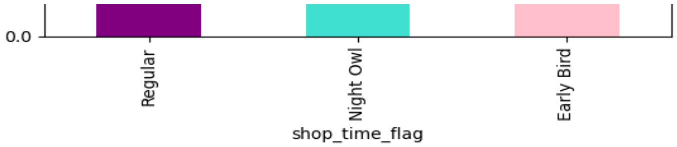
Recommend targeting this group in advertisements.



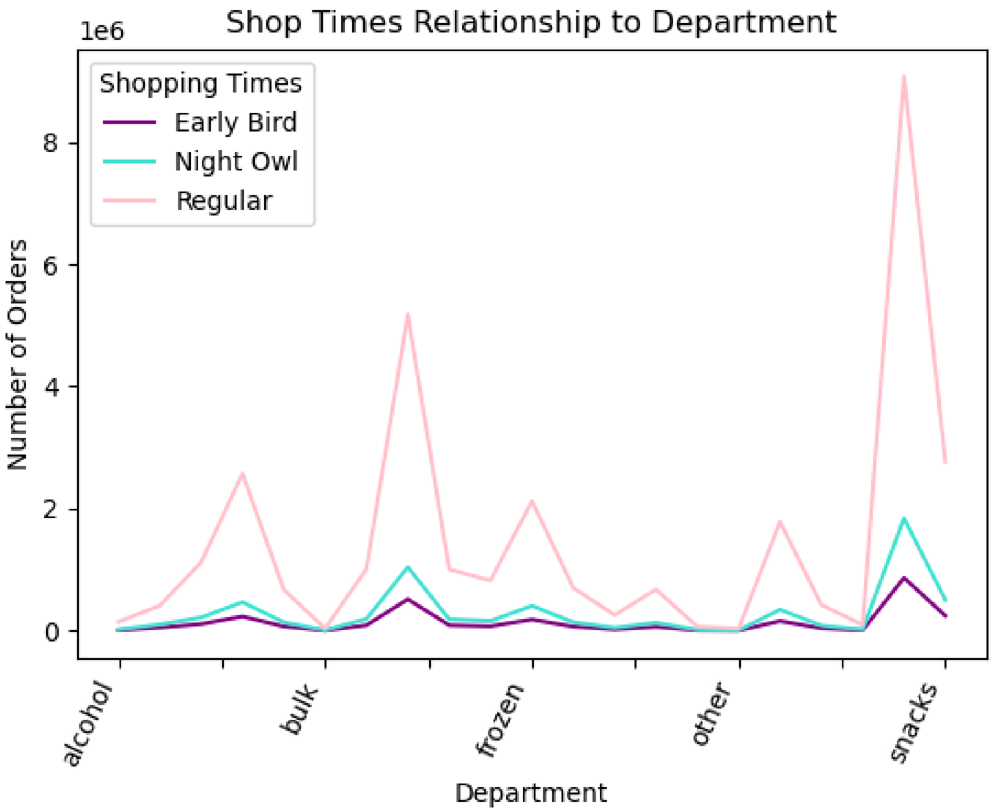
Majority of shopping hours are ocurring during regular daytime hours (80.4% demons chart)







This line chart below shows all shoppers show the same preference patterns as they all buy mor
no matter what time of day the order is placed



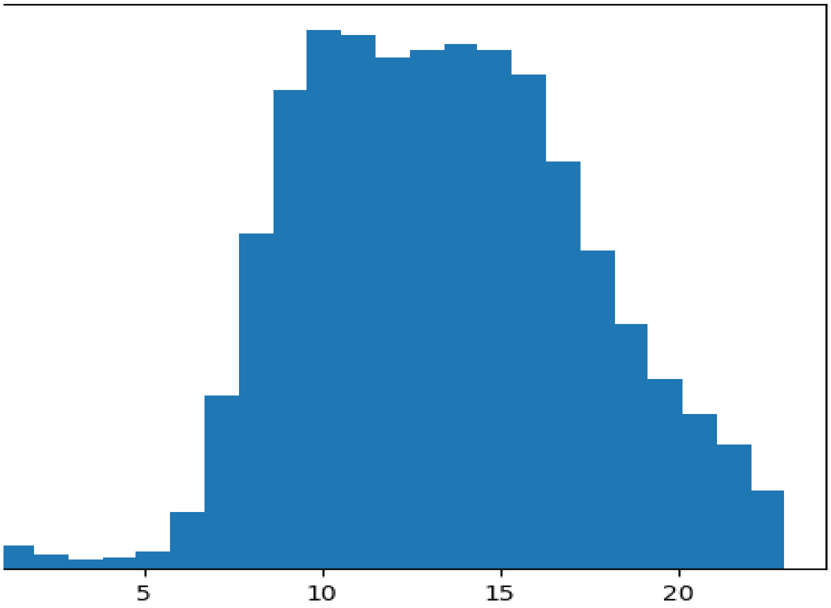
Title page

ɔ with an organized presentation. A few of these visualizations
:ommendations' tab. This page can be referred to for those if

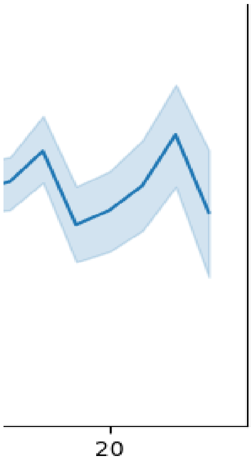
oney, as this might inform the type of products they advertise at these times.

ƷHTS:
he slower times for Instacart. Busiest days are weekends between 12pm-3pm.

ɛrtise conveneint snacks and/or lunch foods as people are most likely busy looking for food 'on the

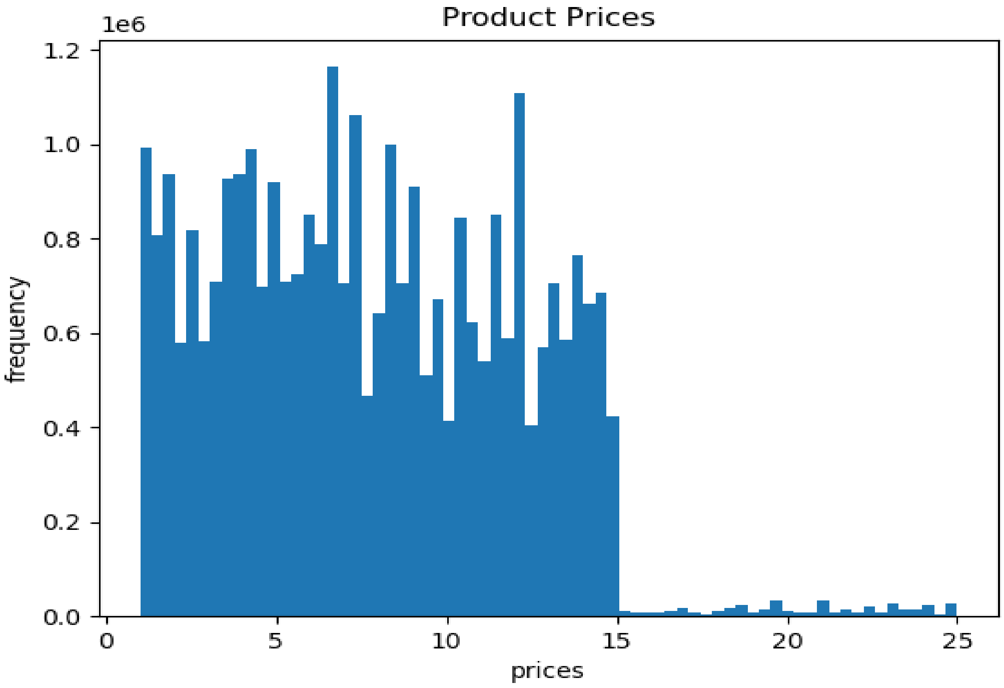


ems are ordered in the very
ɔ people working overnight
ıo are struggling to sleep
teresting to note they have



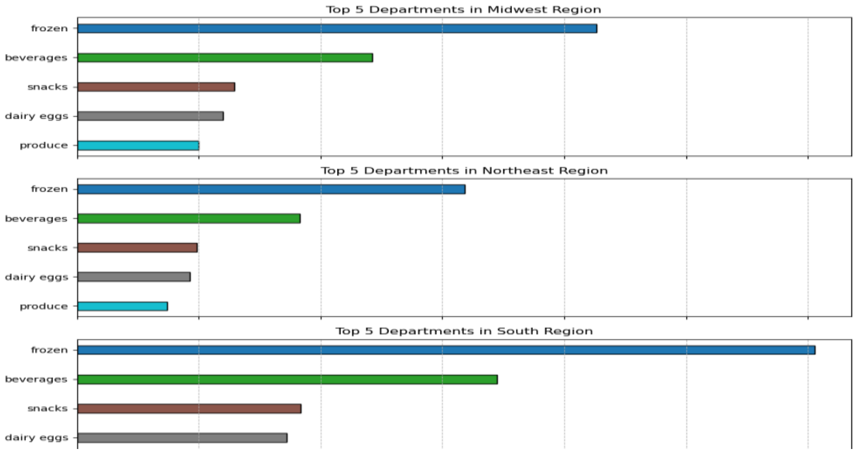
mean_order
verage order
departments
rtments are

INSIGHT:
Most products are between **$1-$15**, while a few are priced higher in $15-$25 range. This information will help to figure out appropriate and simpler price range groupings.



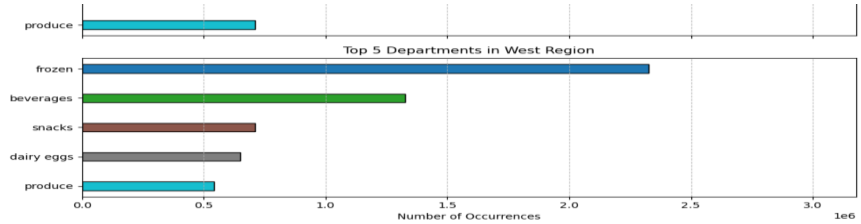
s the top, followed by organic
ar strawberries, limes, and
would be to continue to keep

INSIGHT:
These bar charts are broken up by regions, showing the top 5 departments for that region. All of them show a similar distribution of preference in the following order:
1)**Frozen**, 2) Beverages, 3)Snacks, 4) Dairy Eggs, 5) Produce

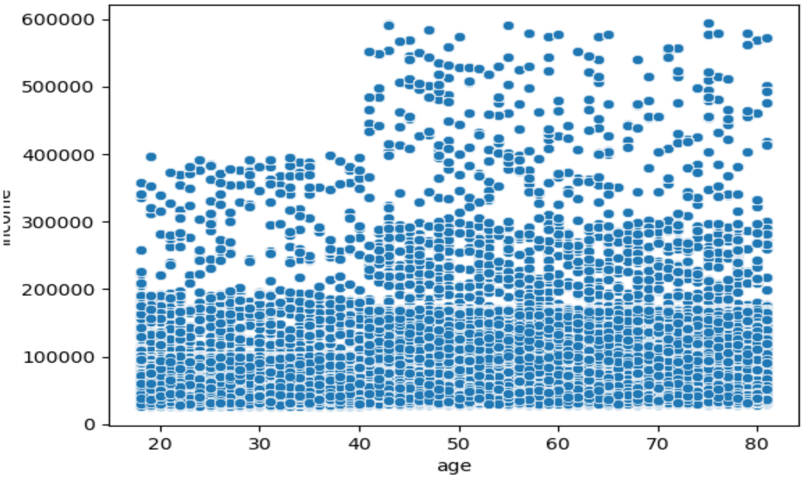

Top 5 Departments in Midwest Region

Top 5 Departments in Northeast Region

Top 5 Departments in South Region

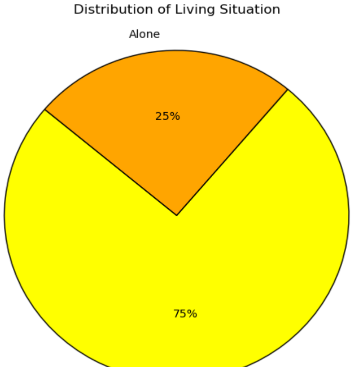


INSIGHT:
lot shows that there is a correlation here between age and spending power represented in plot as there is a **definitive jump in income at age 40**, there is a huge jump from a max of o $600,000. This jump is also represented in what appears to be the average from about 300,000 dollars at age 40 (this information judged by looking at the high density of dots on the scatterplot



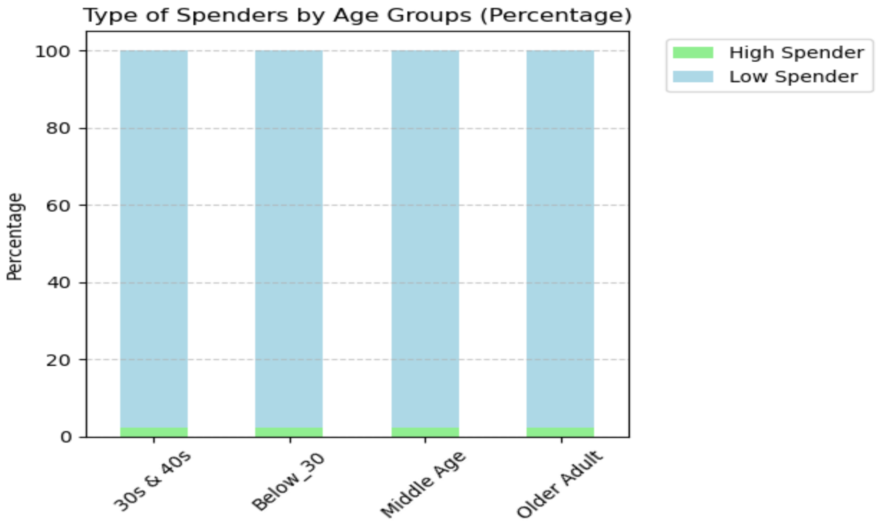
ver

INSIGHT:
Majority of Instacart families have families

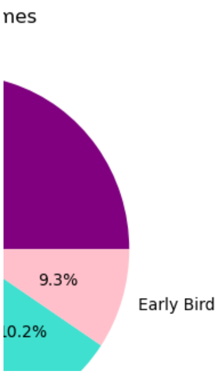


2)
s.

-there is no trend for age in amount spent. According to bar chart below, all age groups seem to be the same 'type' of spender across all age groups (small portion of high spenders and majority of low)



strated in the pie



re snacks,

instacart

Recommendations

Ke

The sales team wants to know what the busiest days of the week and h

Answer:

DAYS:

-Schedule some Ads for slower times during mid-week (Tuesdays & Wedensdays) to promote busine
(since busiest days are on the weekends Saturday (0) and Sunday (1). Test Ads during these slower t
with emphasis on discount items or bundle packages to see if business increases during these times
-Continue Ads during busier weekend times since there is already heavier traffic on the website and
less popular items during these times.

HOURS:

Schedule Ads for mid_day 12pm-3pm, during the busiest hours for Instacart, with emphasis on conv
lunch food. People are most likely busy on their lunch breaks and looking for food to suit their 'on t
promtote bundle packages of items that are becoming stale in inventory in order to get rid of these
-Increase prices slightly (by about50 cents) durin these busier times as people are motivated to mak
mind these small increases.

K

They also want to know whether there are particular times of the day when people

Answer:

The line graph shows that early morning hours are when higher priced items are purchased, arounc

Recommend advertising higher priced items during these hours. Focus on lifestyle or luxury/comfo
be more receptive to ideal living options like backyard BBQ items, candles, kitchen utensils/portable
etc.

K

Instacart has a lot of products with different price tags. Marketing and sales want to use simpler pr
have the highes

Answer:

PRODUCTS:

Produce, dairy/eggs. **Snacks**, **beverages**, **frozen**, and **pantry** departments are the most popular de
The least popular departments are international, alcohol, and pets.

PRICE RANGE:

Most products are between **$1-$15**, while a few are priced higher in **$15-$25** range. This informatic
out appropriate and simpler price range groupings, as majority of items or lower priced. This makes

Recommend placing Ads for coupons on certain items wiithin the popular departments since majori
frequent these areas:
-ie) keeping in mind departments listed above, advertise 3 pack for $10 (discounted from $15) for th
veggies, packs of sodas, yogurts, squeeze apple sauces, and chips.

Also, recommend placing eye-catching, poster Advertisements strategically in these departments as

Are there certain types c

Answer:

The top 10 products ordered: ranked with regular bananas as the top, followed by organic banan
spinach, organic Avocados, lemon, regular strawberries, limes, and organic whole milk.

It is interesting to note the popularity of organic products.

Recommend keeping organic produce well stocked and in good condition to maintain these numl
along with regular promotions of these products. Strategic placement of promotional Ads in ban
products such as organic whole milk, limes, or strawberries

The marketing and sales teams are particularly interested in th

What is the distribution among users in rega
Are there difference
Is there a connection betwe

Answer:
The **30s/40s** age group tends to rank the highest in all aspects of customers according to this bar
They make up 1)the most regular customers
2) the make up the majority of new customers
3) they make up most of the loyal customers

The pie chart shows a general visual of 30s/40s rank as top customers, compared to other age gr
noting that there is not a big variation in the other age groups in relation to eachother.

Majority of Instacart Customers have families.

Recommend targeting these 30s-40s aged, young family groups in advertisements: using young f
appealing to frozen pre-made meals, quick snacks, easy and quick dinner ideas, breakfast on the
likely be drawn to food that fits into their busy lifestyles.

Are there differences in

Answer:

The chart shows that **Southern Regions** generate the most of Instacart's spendings. This is follo
then Midwest Regions, with Northeast spending the least. This pattern is the same for both grou
spenders).

These bar charts are broken up by regions, showing the top 5 departments for that region. Al
distribution of department preference in the following order:
1)**Frozen**, 2) Beverages, 3)Snacks, 4) Dairy Eggs, 5) Produce

Recommend increasing advertisements in southern regions to continue these higher sales, as wel

What different classificat

Answer:

There is a correlation here between age and spending power represented in this scatterplot
in income at age 40, there is a huge jump from a max of $400,000 to $600,000.

This jump in spending power correlates nicely to our representation above showing that 30
customers.

What different classificat
What differences can you find in ordering habits of different customer profiles? Consider th

Answer:

Most orders are placed during regular daytime hours (between 8am and 8pm). Make sure adve
made during these hours. Ads can be cycled throughout the course of the day. For instance, a
during morning hours, lunch foods/snakcs during mid-day hours, and quick dinner options in th
evening.

Try advertising more meal planning options for those in the evenings who prepare for the next
orders are made.

Early Bird shops between 5-8am
Night Owl shops between 8pm-5am
Regular is the default for the other times

The line chart shows that snack items are the most frequently purchased for all shoppers, n
snack options to always be included in advertisements at all hours of the day.

Recommendations Review:

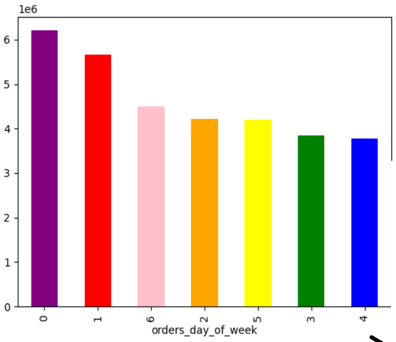
-Increase Ads during mid week (Tuesdays & Wednesdays)
-Schedule Ads between 12pm and 3pm, the busiest hours for p
-Schedule Ads for higher priced items in the early morning hour
-Advertise snacks all hours
-place eye-catching poster Ads within the popular departments
-promotional Ads for popular items to increase amo
-Keep $5 price range groupings as majority of products are betw
-Promotional Ads for organic items (placed in highly populated
-Keep organic foods well stocked and in good shape to maintai
-Recommend targeting 30s-40s, young family groups in adverti
-Recommend increasing advertisements in southern regions to
gain more customers (keeping in mind these variables and targ

Title page

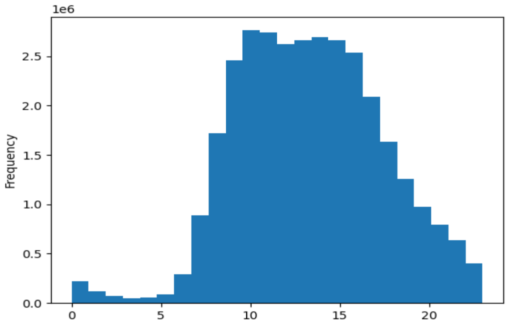
ey Question 1:
hours of the day are in order to schedule ads at times when there are fewer orders.

ess.
times
.
l promote

0 = Saturday
1 = Sunday
2 = Monday
3 = Tuesday
4 = Wednesday
5 = Thursday
6 = Friday



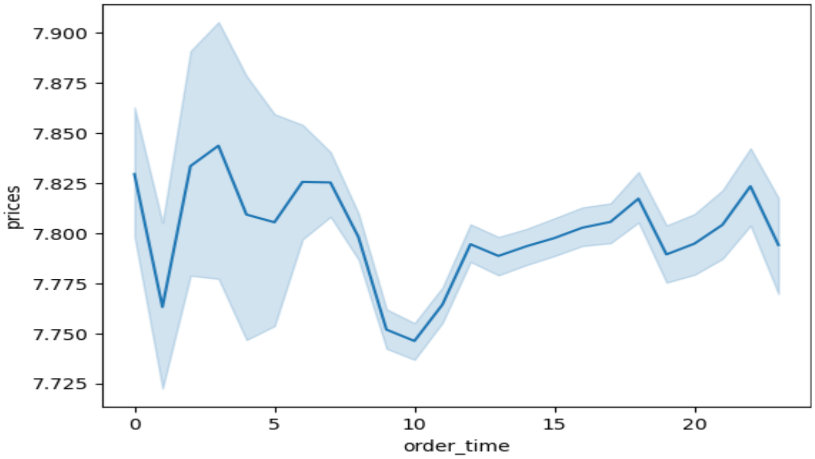
venient snack or pre-packaged
he go' needs for the day. Also,
items while making money.
ke purchases and should not



Key Question 2:
e spend the most moeny, as this might inform the type of products they advertise at these times.

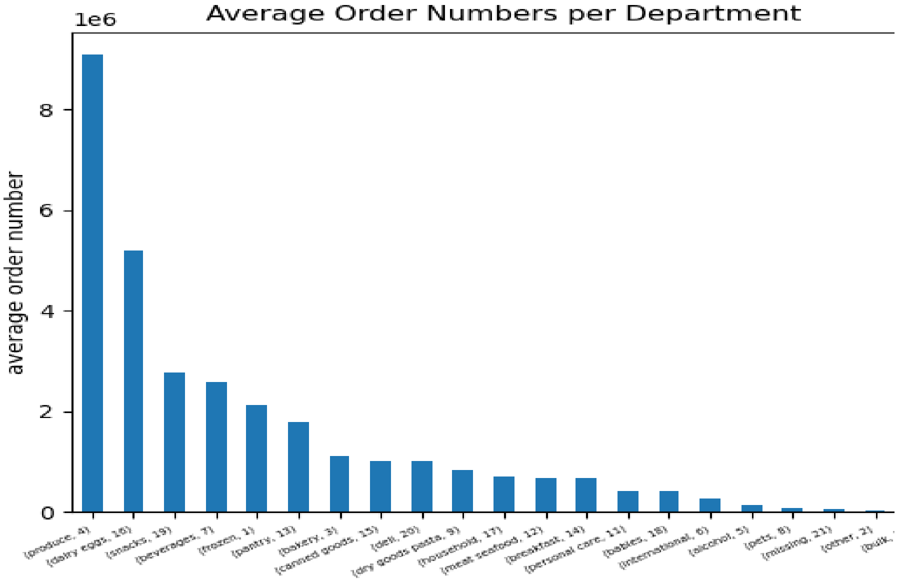
d 2-3am.

ort items , as this population might
e appliances, salon hair products,

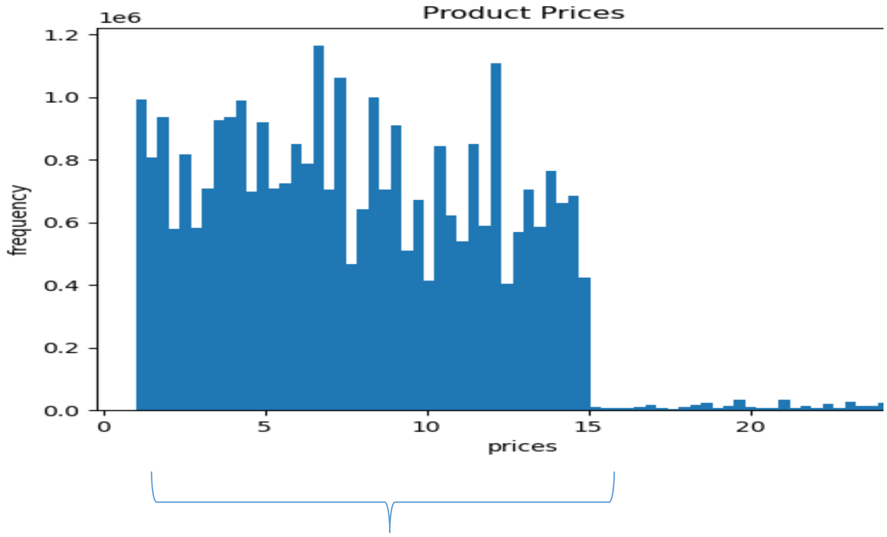


Key Question 3:
rice range groupings to help direct their efforts. The marketing and sales teams want to know which departmen
t frequency of product orders.

partments.



ɔn will help to figure
ꞩ sense for a grocery



ity of shoppers

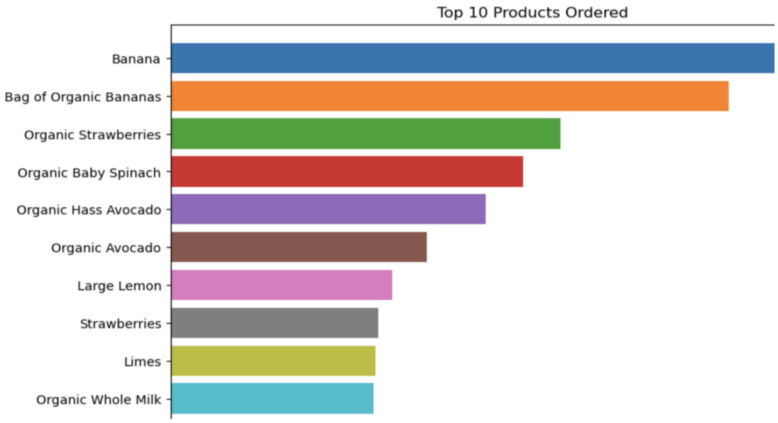
ıings like frozen

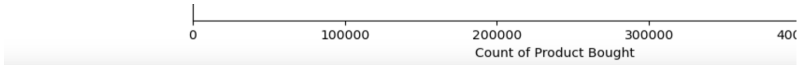
ꞩ well for the lesser

Key Question 4:
ɔf products that are more popular than others?

ıs, organic strawberries, organic baby

bers as well advertising organic products,
ıana section for other less bought





Key Quesiton 5:

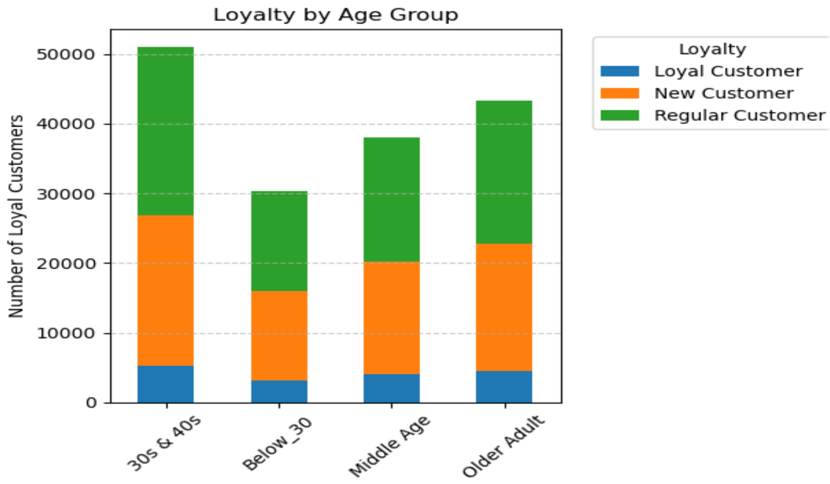
he different types of customers in their system and how their ordering behaviors differ.

ards to their brand loyalty (ie: how often do they return to Instacart?

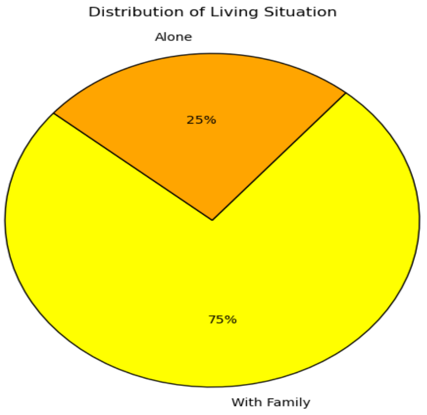
es in ordering habits based on loyalty status?

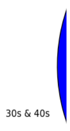
een age and family status in terms of ordering habits?

chart:



oups,



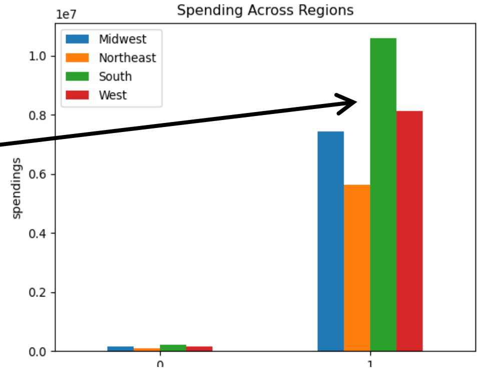


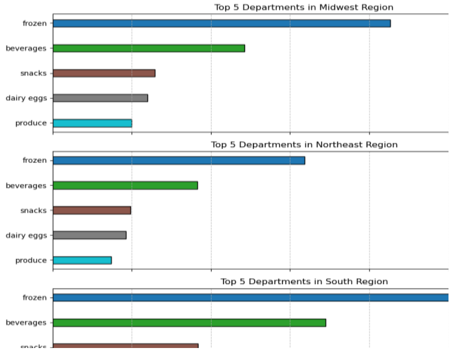
amilies or working adults for the ads, go, etc. This population would most

Key Question 5:

ordering habits based on a customer's region?

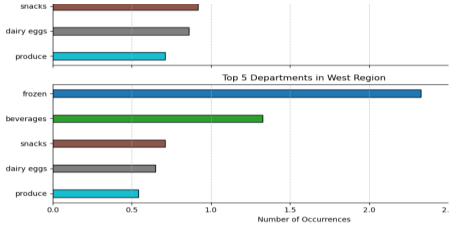
owed by Western Regions, ups of people (high and low





l of them show a similar

ll as promotion of Instacart

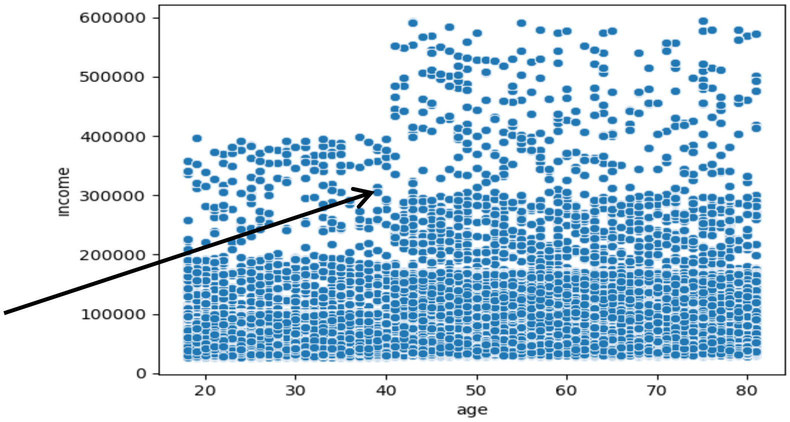


Key Question 6:

tions does the demographic information suggest? Spending power?

t as there is a **definitive jump**

'sand 40s make up majority of



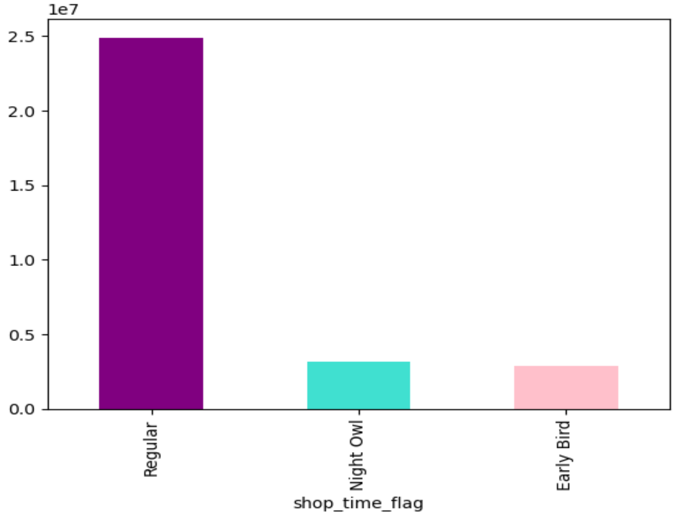
Kep Question 7:

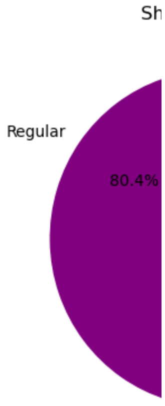
tions does the demographic information suggest?

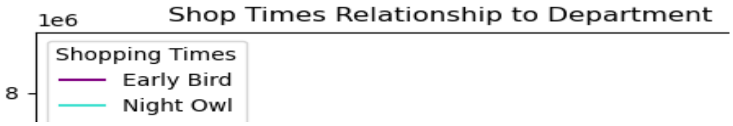
he price of orders, the frequency of orders, the products customers are ordering, and anything else you can thinl

ertisements are predominantly
dvertise quick breakfast items
e evening hours during early

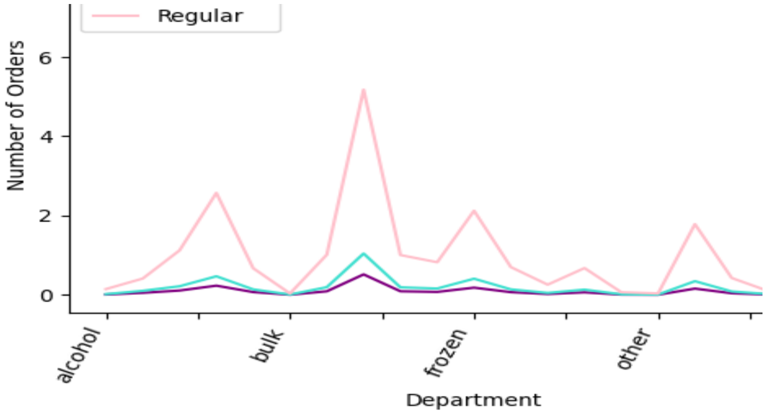
days, prior to 8pm when less







ıo matter the time. Recommend



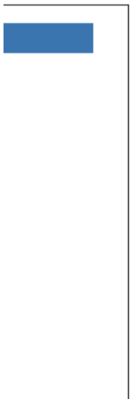
lacing orders.
rs between 2-3am.

(Produce, dairy/eggs. Snacks, beverages, frozen, and pantry)
unt purchased, as well as Ads for lesser populated areas (international, pet, &
ween $5-$15.
banana section)
n the integrity of these sales.
isements - busy lifestyles (prepared food, quick snacks, food on-the-go)
maintain integrity of their higher sales, as well as promotion of Instacart to tl
et populations in the types of Ads used.

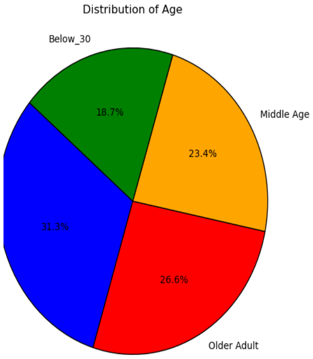
ts

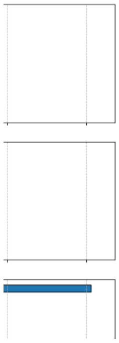
, 10)

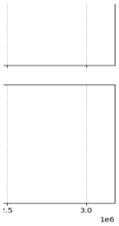
25



0000

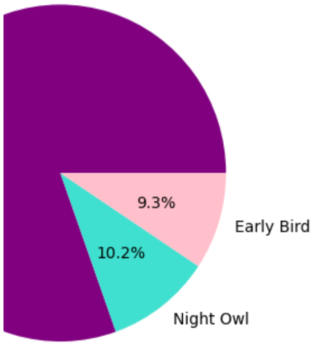

Distribution of Age
Below_30
Middle Age
Older Adult
18.7%
23.4%
26.6%
31.3%





k of.

hopping Times



& alcohol)

he other regions to