Table of Contents

Introduction

A brief description on the pet related goods industry in Hong Kong

Hong Kong, a global city, also is the hub for pet food trade in the region. In recent years, there are more people who like to adopt pets as their new family. To spoil them, pet parents will buy a lot of luxuries and extravagance. Therefore, the pet industry in Hong Kong become stronger. However, why people would like to adopt a pet in such a year, there are some reasons below.

I believe one of the reasons pet related goods industry rise is the demographic changes. In this era, most people are focused on working. It lets Hong Kong facing a decline in the birth rate or getting married later. For that reason, it causes a demographic change in this city. In results, there are many elderlies who are single. To erase their feeling of isolated, adopting a pet is one of the good choices. It makes the pet industry rise. For now, I am sure that why the population are changing the fact of pet related goods industry.

As you can see from the above, most people change their way of raising a pet. The pet owners keep their dogs as another housemate. They love their pets as a member of their family. They want to give them as better as possible. They will buy them different luxuries and services as their lifestyle, such as grooming needs, wellness, and beauty. The businessperson understands pet parents' feelings and develops several types of equipment and services of pet related goods industry in Hong Kong. Pets are important for their pet owners. They treat them as one of family members, they do not step back from costing on the pets. Pet related goods industry are key role in treating a pet. Besides, with more people adopting pets in such years, there is no doubt that the industry is going to increase in the future.

Objectives

In our project, we want to analyze different data to find the behavior of most customers when they choose to buy a product. Therefore, our objectives are related to the product such as **product selection in the store**, **product quality** and **product price**. We will analyze those data to output a lot of conclusions. About the qualitative data, we will use pie chart to show which type of goods usually buy, use the Cross tabulation to analyze the relationship between type of goods usually buy and maximum time to visit the pet good shop in a month and use cross tabulation to find out which type of goods have been bought by each gender. For the Quantitative data, we will use different columns of data to figure out the product quality, product selection in the store and product price by the method of Summary Statistics or Scatter Diagram. Eventually, we can figure out the behavior of customers in our shop. Therefore, we can according from the results to make our shop become better, provide the good products or services to let their pet become more comfortable in their life.

Data Processing and Cleaning

Problem	Spelling mistakes	
Solution	Replace with the most data option in the same column with the same meaning	
Column	A, B, D, AQ	
Column Example	A	Change "1 wee before" to "1 week before" and "Last mouth" to "Last month"
	B	Change "To" to "Toy" and "Snake" to "Snack"
	D	Change "E-M" "E-MoM 汴" to "E-Money" and "Cas" to "Cash"
	AQ	Change "Mal" to "Male" and "Man" to "Men"

| Problem	Having Synonyms	
Solution	Replace with the most data option in the same column with the same meaning	
Column	A, B, D, AQ, AR, AS, AU, AX	
Column Example	A	Change "No" "Nil" to "No visit within last 6 months"
	B	Change "Any" to "Random", "Pet Toy" "Play" "玩具" to "Toy" and "Training Snack" to "Snack"
	D	Change "Visa" to "Credit Card"
	AQ	Change "Women" "Woman" "F" to "Female" and "M" "Men" to "Male"
	AR	Change "40+" to "40 or above"
	AS	Change "F.7" "F5" "Form 5" "Form 7" "Grade 12" "Secondary" to "Senior High School"
	AU	Change "Facebook" "IG" "Instagram" to "Social Media" ,"Mum" "Sibling" to "Relative" and "朋友" to "Friends"
	AX	Change "N/A" "Blank" to "0"

Problem	Intentional misreporting
	(Column C) Maximum time to visit the pets good shop in a month must not be 0 if (Column A) last time visiting the pet goods shop is this week or last week
Solution	delete the entire row that answers 0 in column C and week in column A
Deleted row	3 rows deleted and number of rows changed from 1016 to 1013

Problem	Missing data	
Solution	Delete the record with missing data if the entire row of data in column E to AP is missing Mean substitution if there are less than 3 missing data in a row	
Column	E to AP	
Column Example	E, F, N to T, AE, AF, AN	Sub 3 to the blank cell
	G to M, U to AD, AG to AM, AO, AP	Sub 4 to the blank cell
Deleted row	6 rows deleted and number of rows changed from 1013 to 1007	

Problem	Blank cell	
Solution	Delete the record with missing data	
Column	AQ, AU	
Column Example	AQ	1 row deleted
	AU	3 rows deleted

Problem	BBA is not an education level	
Solution	Delete the record with missing data	
Column Example	AS	3 rows deleted

Prove that the values of column AT are normally distributed
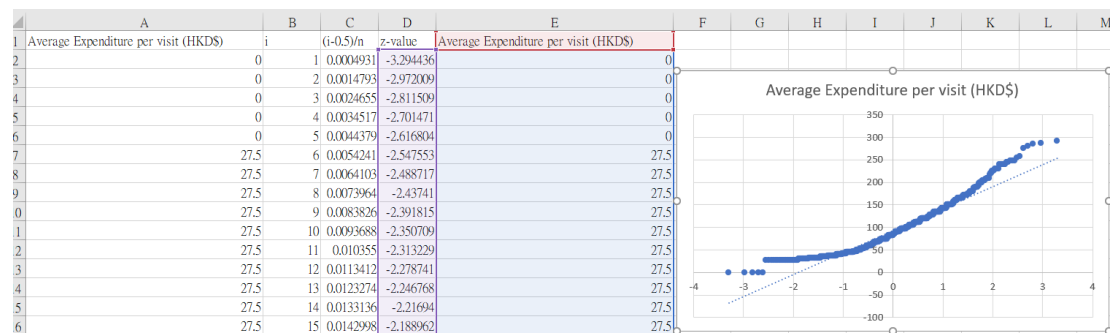
After deleting the value "25000", Average Expenditure per visit (HKD$) is right-skewed distribution.

Column	AT
Problem	**Outliers**
Solution	After deleting the row with the value of 25000, following rules of Thumb, the value range should be below 244.5. Since the value is related to money, the value should be 0 or above.
Example	Change values above 244 to 244

Column	AV	AX	AW
Problem	**Extreme value**		
Solution	Replace numbers greater than 100 with 100 and negative numbers with 0, because when the unit is a percentage, the maximum number should be 100 and the minimum number should be 0		Delete the record with cell is 0 because it is impossible to leave within 1 minute
Column Example	AV	Change "116" "112" "108" "104" to "100"	
	AX	Change "-3" "-0.3" to "0"	

After cleaning the data, a total of 22 rows were deleted and leaving 994 rows.
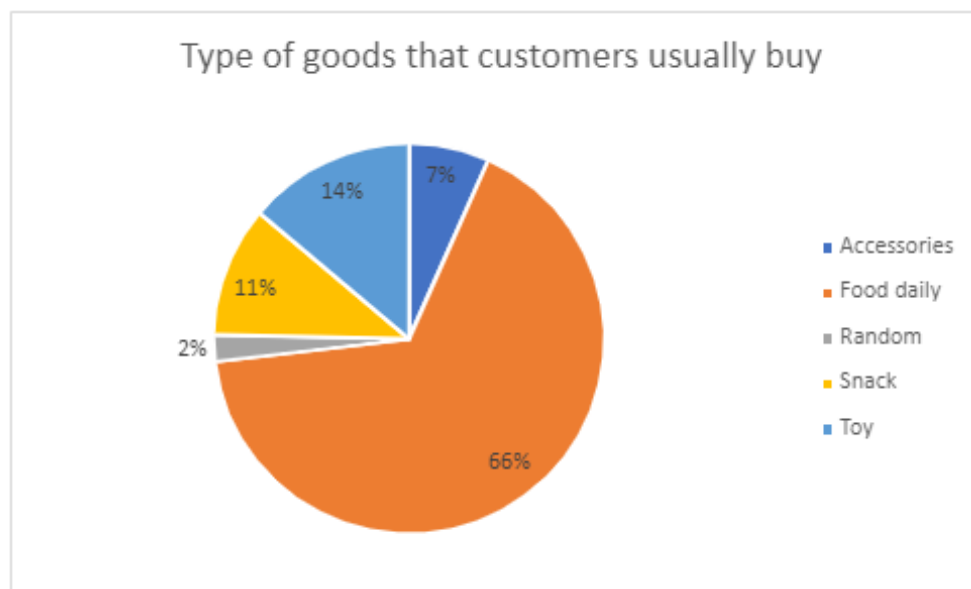
Descriptive Analysis

Qualitative

Objective (Product selection in the store)

Method: Pie Chart
Columns: Type of goods usually buy

The above pie chart shows the types of goods that customers usually buy. We can see that two-thirds of all the respondents buy daily food for their pets. The second type of good that customers usually buy is toys. The percentage is 14% of all the respondents. Snacks are in third place according to the pie chart. Around 11% of respondents usually buy snacks for their pets. Lastly, less than 10% of the respondents buy accessories and only 2% of the respondents will buy different goods randomly. Based on the pie chart, we can conclude that daily food is the most attractive goods in the store because many respondents usually buy daily food in the store.

Method: Cross tabulation

Columns: Type of goods usually buy, Gender

Count of Type of goods usually buy	Gender		
Type of goods	Female	Male	Grand Total
Accessories	37	29	66
Food (Daily)	367	293	660
Random	9	13	22
Snack	64	43	107
Toy	74	64	138
Grand Total	551	442	993

The above cross-tabulation shows the attractiveness of goods to different genders. The higher the number of goods sold to the customers, the more attractive the goods are. The result shows that the most attractive for females and males are the daily food. It covers 66.6% of the females and 66.3% of the males. The remaining are toys, snacks, and accessories according to the number of goods sold to customers. For toys, it covers 13.4% of the females and 14.5% of the males. For Snack, it covers 11.6% of the females and 9.7% of the males. Lastly, for accessories, it covers 6.7% of the females and 6.6% of the males. Based on the cross-tabulation, we can conclude that both females and males usually buy daily food for their pets. Toys are the second most goods usually bought, snacks are the third, and the accessories are located on the fourth place.

Method: Cross tabulation

Columns: Type of goods usually buy Maximum time to visit the pets good shop in a month

Count of Type of goods usually buy	Maximum time (hours)								
Type of goods		0	1	2	3	4	5	6	Grand Total
Accessories		0	40	14	5	1	5	1	66
Food (Daily)		8	233	163	153	57	34	12	660
Random		0	16	6	0	0	0	0	22
Snack		1	45	32	17	5	5	2	107
Toy		1	73	29	24	6	4	1	138
Grand Total		10	407	244	199	69	48	16	993

The above cross-tabulation shows the relationship between the type of goods usually bought and the maximum time of shopping in the store. 41.0% of customers shop for a maximum time of 1 hour in the store, 24.6% of customers shop for 2 hours in the store, 20.0% of customers shop for 3 hours, 6.9% of customers shop for 4 hours, 4.8% of customers shop for 5 hours, and lastly, 1.6% of customers shop for 6 hours. For any hours in the cross-tabulation, daily food is the most popular goods in the store. The percentage of customers buying daily food to the number of customers which maximum time of visiting the shop is from 0 - 6 hours are 80%, 57.2%, 66.8%, 76.9%, 82.6%, 70.8%, and 75% respectively. On the other hand, accessories are the least favorite good for all the respondents. The percentage of customers buying accessories to the number of customers which maximum time of visiting the shop is from 0 – 6 hours are 0%, 9.8%, 2.7%, 2.5%, 1.4%, 10.4%, and 6.3% respectively. Based on this cross-tabulation, we can conclude that daily food can attract customers to visit the store longer while other products can only attract most of the customers to visit the store within the first few hours.

Quantitative

Objective (Product quality)

Method: Summary Statistics
Columns: Quality (Importance), Quality (Satisfaction Level), Freshness (Importance), Hygiene (Importance), Hygiene (Satisfaction Level), Freshness (Satisfaction Level)

	Quality (Importance)	Quality (Satisfaction Level)	Hygiene (Importance)	Hygiene (Satisfaction Level)	Freshness (Importance)	Freshness (Satisfaction Level)
Mean	3.464249748	3.435045317	3.50755287	3.347432024	3.996978852	3.694864048
Standard Error	0.03139604	0.030793356	0.032744718	0.028977305	0.031589612	0.034881608
Median	3	4	4	3	4	4
Mode	3	4	4	3	4	4
Standard Deviation	0.989348956	0.970357237	1.031848362	0.913130003	0.995448785	1.099185827
Sample Variance	0.978811357	0.941593168	1.064711042	0.833806403	0.990918283	1.208209483
Kurtosis	0.072417199	0.123968903	-0.059398717	0.193961686	1.239737764	-0.219378659
Skewness	-0.42003701	-0.648764765	-0.467120559	-0.311458175	-1.124463516	-0.661433234
Range	4	4	4	4	4	4
Minimum	1	1	1	1	1	1
Maximum	5	5	5	5	5	5
Sum	3440	3411	3483	3324	3969	3669
Count	993	993	993	993	993	993

The above table shows the summary statistics of quality, hygiene, and freshness' importance and satisfaction level rated by the customers. In general, the mean of these three types of satisfaction levels is all lower than the importance. It shows that the customers may not be satisfied with the product's quality, hygiene, and freshness. Also, all fields' standard deviation is between 0.9 and 1.1. It shows that the customers' responses are quite stable. Also, the mean of freshness importance and satisfaction level are both the highest. It shows that customers are placing emphasis on the product's freshness. But at the same time, even if customers are satisfied with product freshness, they may not be satisfied with product quality because the quality satisfaction level is lower than the freshness satisfaction level. In addition, the mean of Quality (Satisfaction level) is similar to Hygiene (Satisfaction level). It shows that customers who are satisfied with product quality are also satisfied with product hygiene. On quality (Satisfaction Level), its skewness is -0.648, which is < 0, and its kurtosis is 0.124, which is > 0. This means it is a left-skewed heavy-tailed distribution. On quality (Importance), its skewness is -0.420, which is < 0, and its kurtosis is 0.072, which is > 0. This means it is a left-skewed heavy-tailed distribution. On Hygiene (Satisfaction Level), its skewness is -0.311, which is < 0, and its kurtosis is 0.194, which is > 0. This means it is a left-skewed heavy-tailed distribution. On Hygiene (Importance), its skewness is -0.467, which is < 0, and its kurtosis is -0.059, which is < 0. This means it is a left-skewed light-tailed distribution. On Freshness (Satisfaction Level), its skewness is -0.661, which is < 0, and its kurtosis is -0.219, which is < 0. This means it is

a left-skewed light-tailed distribution. On Freshness (Importance), its skewness is -1.124, which is < 0, and its kurtosis is 1.240, which is > 0. This means it is a left-skewed heavy-tailed distribution.

Objective (Product selection in the store)

Method: Summary Statistics

Columns: Choice (Satisfaction Level), New Product (Satisfaction Level)

	Choice (Satisfaction Level)	New Product (Satisfaction Level)
Mean	3.793762575	3.284708249
Standar Error	0.033575724	0.032931448
Median	4	3
Mode	4	3
Standard Devistion	1.058567563	1.038254968
Sample Variance	1.120565285	1.077973379
Kurtosis	0.236100137	-0.217659316
Skewness	-0.770574445	-0.178713427
Range	4	4
Minimum	1	1
Maximum	5	5
Sum	3771	3265
Count	994	994

The above table shows the summary statistics between Choice and the new product rate for satisfaction level by the customers. First, we find the mean and we can see that choice's mean is larger than new product's mean. We know that our customers may think our choice is enough and they are satisfied but it is not enough satisfaction for our new product. Their skewness is a negative number. Therefore, we know that they are both left skewed distribution. Next, in the Choice for satisfaction level is a heavy tailed distribution because we can see that there is a positive kurtosis. Then, in the New Product for satisfaction level is a light tailed distribution since we can see that there is a negative kurtosis.

Method: Summary Statistics

Columns: Choice (Importance), New Product (Importance)

	Choice (Importance)	New Product (Importance)
Mean	3.895372233	3.277665996
Standar Error	0.035914236	0.033615574
Median	4	3
Mode	5	3
Standard Devistion	1.132295618	1.059823935
Sample Variance	1.282093366	1.123226773
Kurtosis	0.373027955	-0.2696308
Skewness	-1.011541802	-0.205453063
Range	4	4
Minimum	1	1
Maximum	5	5
Sum	3872	3258
Count	994	994

The above table shows the summary statistics between Choice and the new product rate for importance by the customers. First, we find the mean and we can see that choice's mean is larger than new product's mean. We know that our customers may think our number of choices is importance than the new product. Their skewness are also a negative number. Therefore, we know that they are both left skewed distribution. Next, in the Choice for importance is a heavy tailed distribution because we can see that there is a positive kurtosis. Then, in the New Product for satisfaction level is a light tailed distribution since we can see that there is a negative kurtosis.

Objective (Product price)

Method: Scatter Diagram and r
Columns: Price (Importance) (y), Quality (Importance) (x)

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.372724131
R Square	0.138923278
Adjusted R Square	0.138054381
Standard Error	1.025100657
Observations	993

r = 0.3727
There is a weak positive linear correlation between the rating on Price (Importance) and rating on Quality (Importance)

Scatter Diagram

When Quality (Importance) increase, Price (Importance) also increase.

It can be seen that most customers believe that the quality of the product meets their expectations for the price of the product

Method: Scatter Diagram and r
Columns: Price (Satisfaction Level) (y), Quality (Satisfaction Level) (x)

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.48950633
R Square	0.239616447
Adjusted R Square	0.238849158
Standard Error	0.89634245
Observations	993

r = 0.4895

There is a weak positive linear correlation between the rating on Price (Satisfaction Level) and rating on Quality (Satisfaction Level))

Scatter Diagram

When Quality (Satisfaction Level) increase, Price (Satisfaction Level) also increase.

It can be seen that most customers who think that product quality is very important think that product price is also an important factor when considering purchasing a product.