SMDM Project

Data Analysis Report

Prepared By

JAI GOUTHAM V

Submitted on
13-09-2020

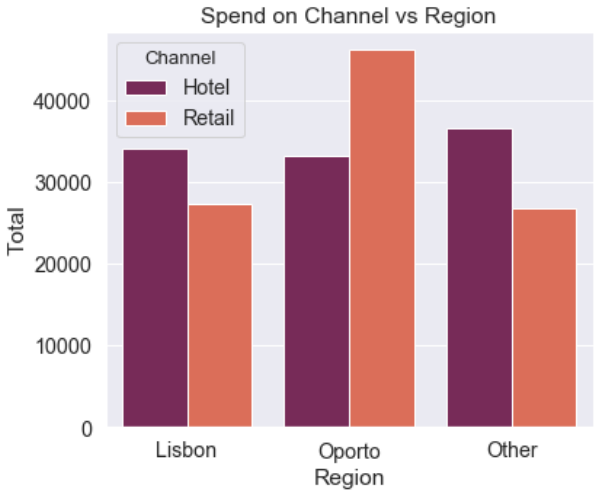
Problem 1:

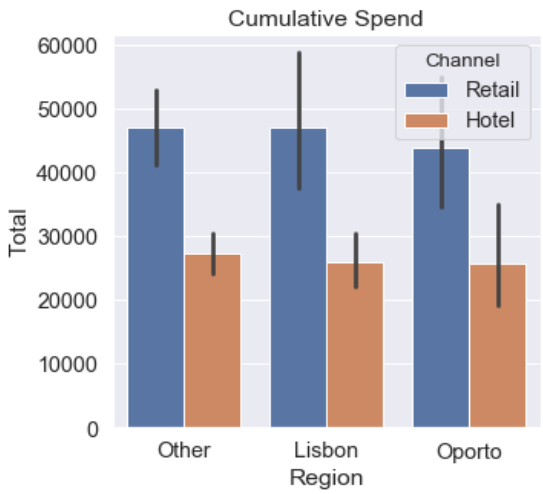
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Before getting into the problem the basic exploratory data analysis can be done in data frame:

- Find the shape of the data, data type of individual columns
- Check the presence of missing values
- Descriptive stats of numerical columns and distribution of skewness and outliers.

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?



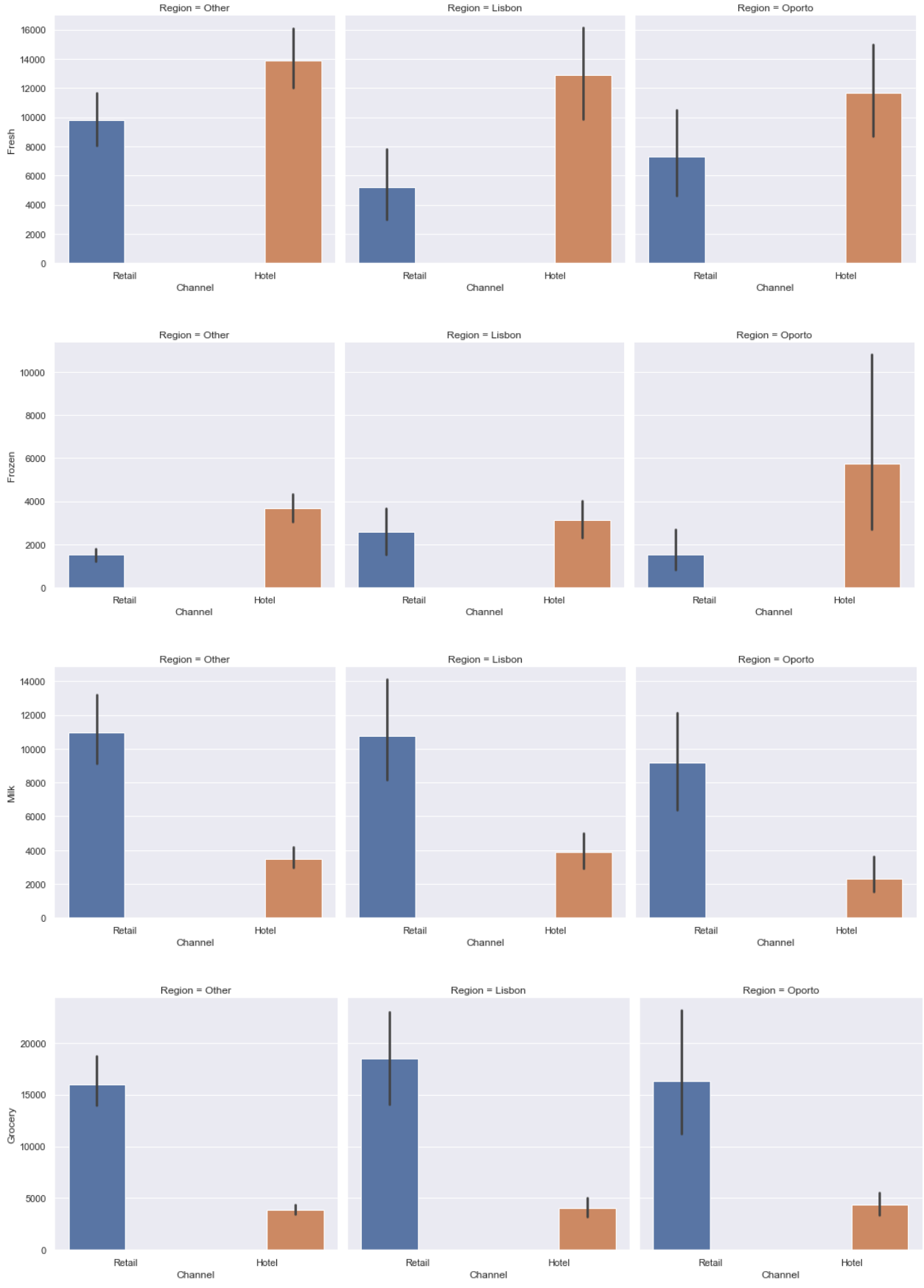


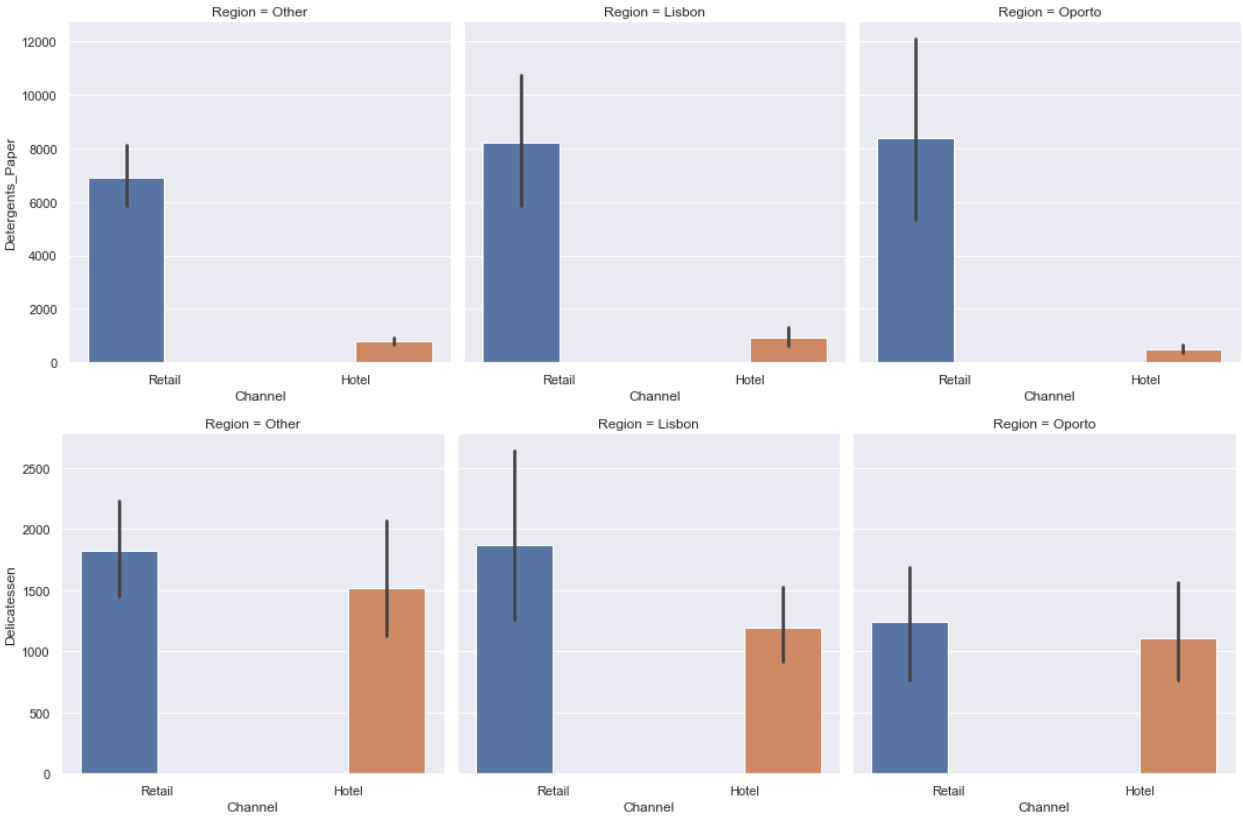
- In **Channel – Retail** and **Region-Oporto** spends **more** in all areas like Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen.

- In **Channel – Retail** and **Region-Other** spends **less** in all areas like Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen.

1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?

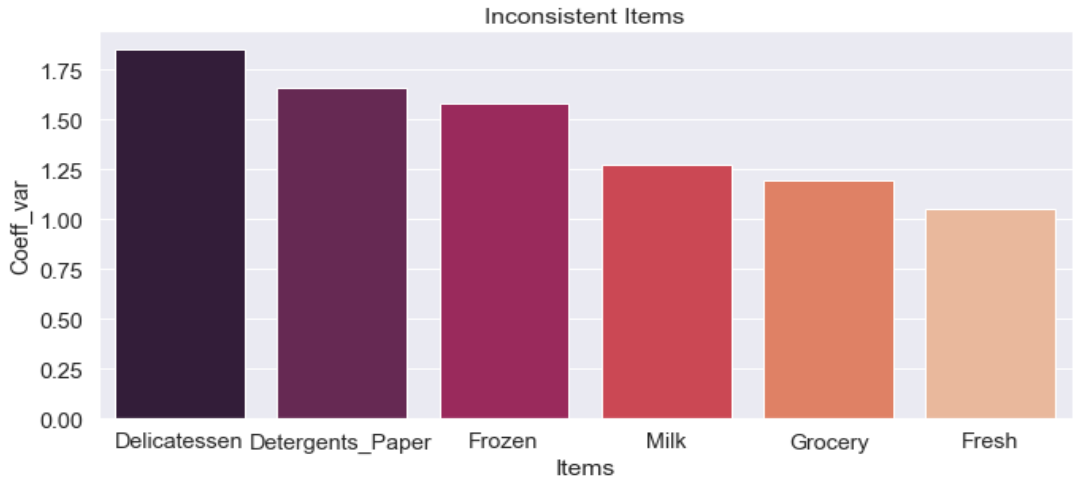
	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0


Region = Other
Region = Lisbon
Region = Oporto
Fresh
Frozen
Milk
Grocery
Channel
Retail
Hotel
0
2000
4000
6000
8000
10000
12000
14000
15000
16000
20000



- The **minimum** spend for items like Fresh, Grocery, Detergents paper, Delicatessen shows similar behavior.
- The **variation** of spend on items like Frozen and Detergents paper shows similar behavior.
- The **average** spend on Frozen and Detergents paper shows almost similar behavior.

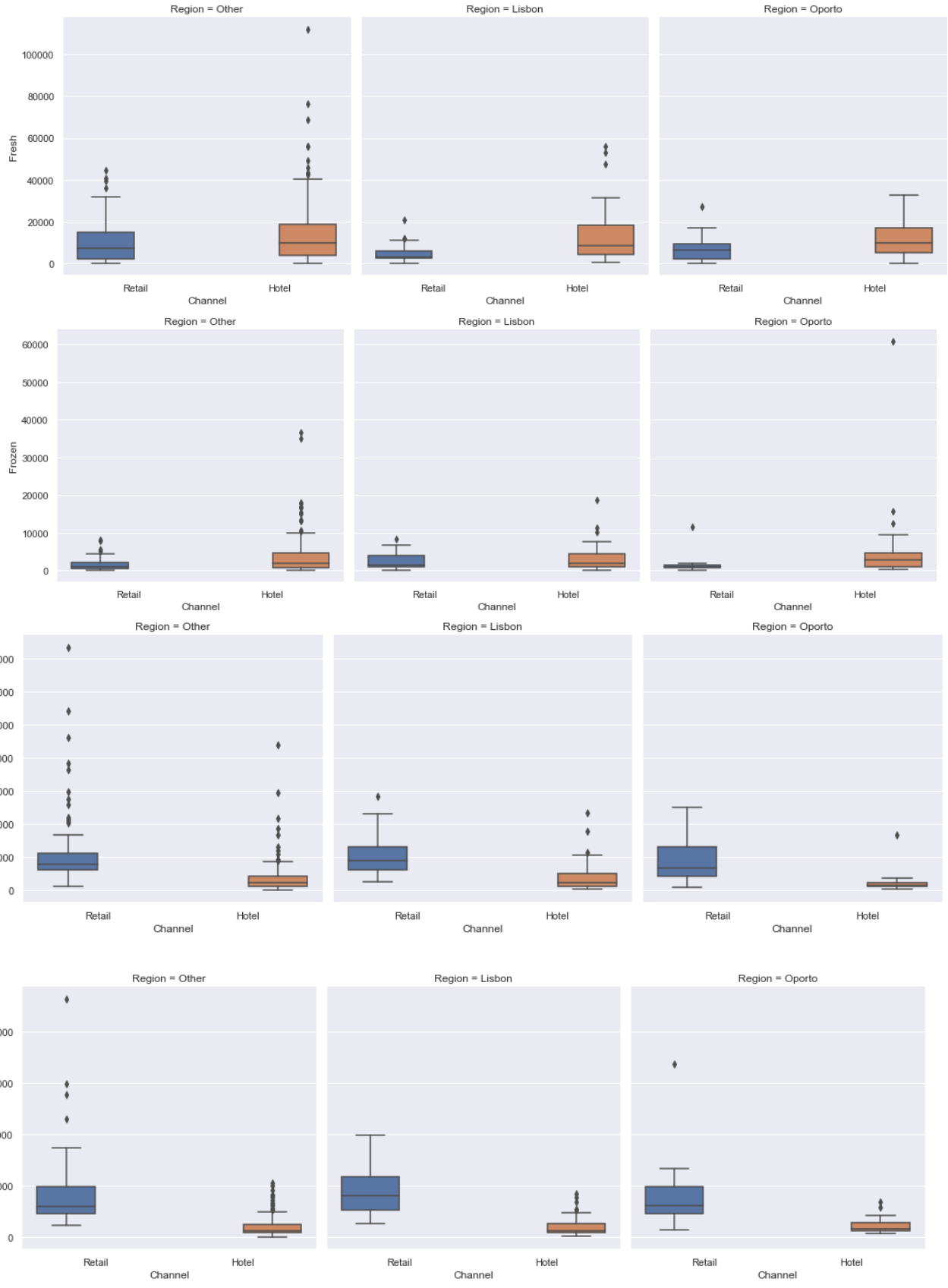
1.3 On the basis of descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

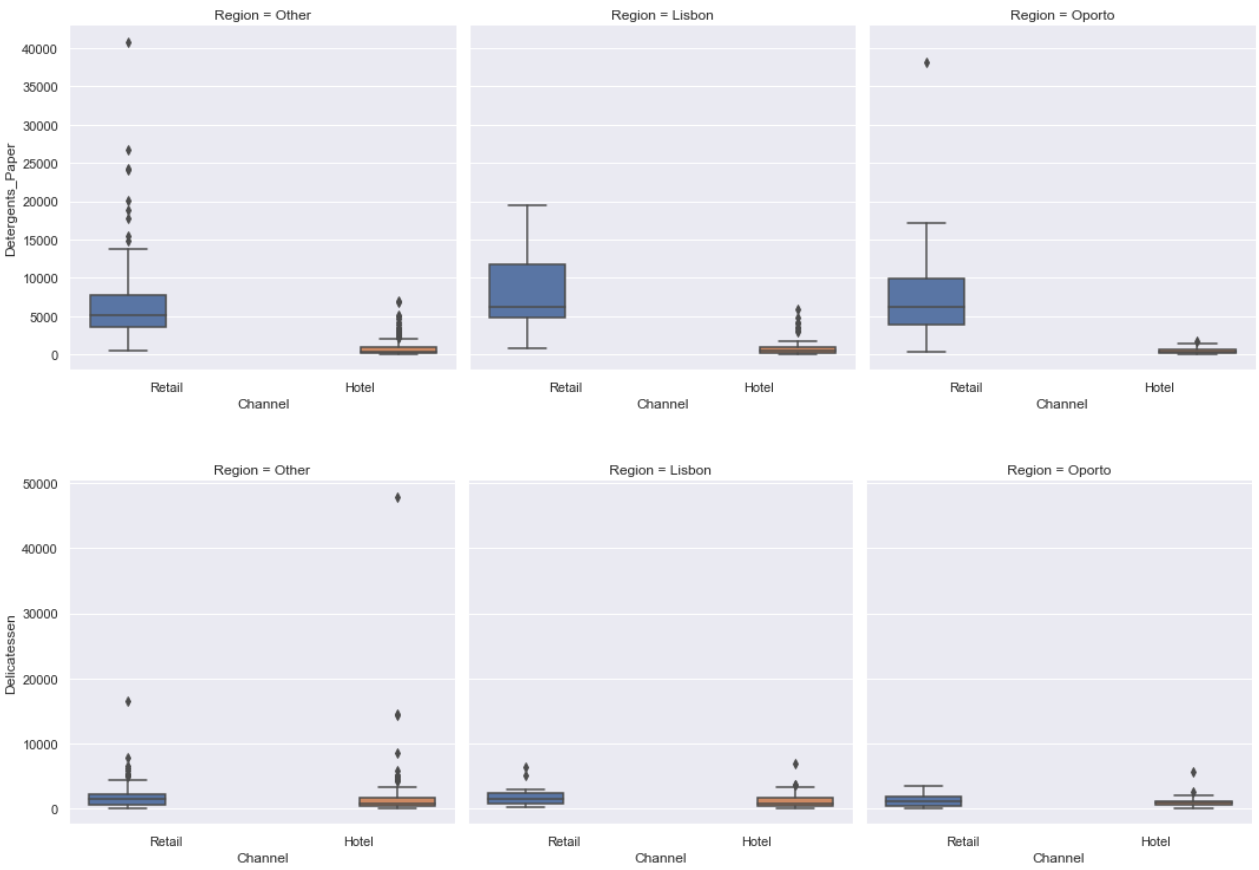


From the coefficient of variation, we can able to find above insights:

- The **most** inconsistent item is **Delicatessen.**
- The **least** inconsistent item is **Fresh.**

- **1.4 Are there any outliers in the data?**





Yes, the data contains outliers in it.

The data which does not contain outliers are given below: -

- The item 'Delicatessen' in Region=Oporto and Channel=Retail which has **not** contain any outliers
- The item 'Detergent Paper' in Region=Lisbon and Channel=Retail which has **not** contain any outliers
- The item 'Grocery' in Region=Lisbon and Channel=Retail which has **not** contain any outliers
- The item 'Milk' in Region=Oporto and Channel=Retail which has **not** contain any outliers
- The item 'Fresh' in Region=Oporto and Channel=Hotel which has **not** contain any outliers

1.4 On the basis of this report, what are the recommendations?

- In **Channel – Retail and Region -Oporto** Necessary measures has to be taken to reduce the expenditure for the items like Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen.
- In **Channel - Hotel and Region -Other** is spending less on items. The same Spending profile/ways can be applied on Channel-Retail and Region-Others to avoid more expenses.
- There is More risk (more inconsistent) in spending on the item **Delicatessen,** the amount spending on this item is varies significantly
- There is least risk (less inconsistent) in spending on the item **Fresh,** there is a consistent spending behaviour on this item.

Problem 2: Probability and distribution

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the **Survey** data set).

2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

2.1.2. Gender and Grad Intention

Grad Intention	No	Undecided	Yes	All
Gender				
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

2.1.3. Gender and Employment

Employment	Full-Time	Part-Time	Unemployed	All
Gender				
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

2.1.4. Gender and Computer

Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

2.2. Assume that the sample is representative of the population of CMSU.

2.2.1. What is the probability that a randomly selected CMSU student will be male?

The number of male students is **29**

The total students are students are **62**

The probability of randomly selected CMSU student will be a male is **46.77%**

2.2.2. What is the probability that a randomly selected CMSU student will be female?

The number of female students is **33.**

The total students are students are **62**

The probability of randomly selected CMSU student will be a Female is **53.22%**

2.3.1. Find the conditional probability of different majors among the male students in CMSU.

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

The conditional probability for male student in managements is **20.69 %**

The conditional probability for male student in Retail is **17.24 %**

The conditional probability for male student in Economics is **13.79 %**

The conditional probability for male student in Accounting is **13.79 %**

The conditional probability for male student in Others is **13.79 %**

The conditional probability for male student in International Business is **6.8 %**

The conditional probability for male student in CIS is **3.4 %**

The conditional probability for male student in undecided is **10.34 %**

2.3.2 Find the conditional probability of different majors among the female students of CMSU.

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

The conditional probability for female student in managements is **12.12 %**

The conditional probability for female student in Retail is **27.27 %**

The conditional probability for female student in Economics is **21.21 %**

The conditional probability for female student in Accounting is **9.09 %**

The conditional probability for female student in Others is **9.09 %**

The conditional probability for female student in International Business is **12.12 %**

The conditional probability for female student in CIS is **9.09 %**

The conditional probability for female student in undecided is **0 %**

2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.

The number of male students, intends to graduate is **17**

The total students are students are **62**

The probability of randomly selected CMSU student will be a male and intends to graduate is **27.41%**

2.4.2 Find the prob. that a randomly selected student is a female and does NOT have a laptop.

The number of female students and does not have laptop are **4**

The total students are students are **62**

The probability of randomly selected student will be a female and does not have laptop is **6.45%**

2.5. Assume that the sample is representative of the population of CMSU, Answer the following :

2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment?

The number of male students is **29**

The total students are students are in fulltime job is **10**

The student who is male and fulltime is **7**

The probability of randomly selected CMSU student will be a male or has fulltime employ is **27.41%**

2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

The conditional probability of randomly selected female student, she is major in international business or management is **24.24%**

2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17
All	12	28

No, the graduate intention and being female are not independent events. Because they are mutually exclusive events.

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

The probability of randomly selected student that his/her GPA is less than 3 is **6.45%**

2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0	All
Gender																				
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1	33
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1	29
All	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

The conditional probability that a randomly selected male earns 50 or more is **48.27 %**

The conditional probability that a randomly selected female earns 50 or more is **54.54 %**

2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

From the above Histogram and Box plot we can conclude the following statements:

- The **GPA** which shows normal distribution, has no outliers.
- The **Salary** shows slightly normal distribution and it has outliers.
- The **Spending** right skewed distribution and it has outliers.
- The **Text messages** right skewed distribution and it has outliers.

Problem 3:

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests.

A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet are calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles

Solution:

3.1 Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

1.Hypothesis formulation:

Decide null and alternate hypothesis

HO: μ = 0.35

Null Hypothesis Ho states that the mean moisture content is equal to permissible limit for 0.35 pound per 100 square feet

Ha: μ < 0.35

Alternate Hypothesis Haa states that the mean moisture content is less than the permissible limit for 0.35 pound per 100 square feet

2. Decide the level of significance:

The level of significance is not mentioned in problem statement, so generally we consider 5%

α = 0.05

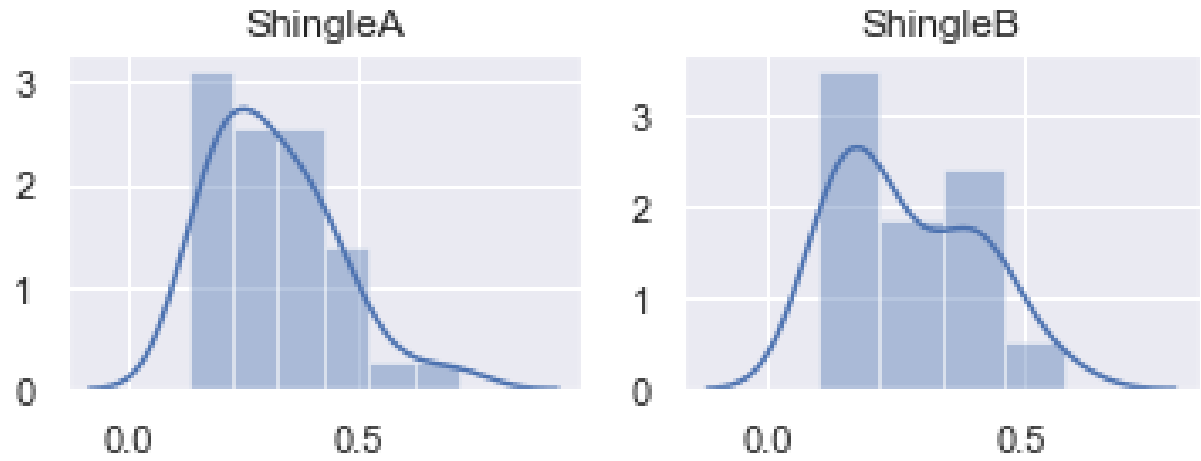
3. Assumptions:

Test of Means using two methods:

3.1) Histogram (visual Test)

3.2) Shapiro test (Statistical Test)

3.1 Histogram:



From the above histogram, The Shingle A - visually the data looks normally distributed.

The Shingle B - visually the data looks like slightly right skewed distribution

3.1 Shapiro Test:

The shapiro test-statistic for sample A (A Shingle) is: 0.9375
The shapiro test P-value for sample A (A Shingle) is: 0.0426

(Ref: Python code data file)

From the Shapiro test, as per our assumptions for Sample A the P-value < alpha.

The shapiro test-statistic for sample B (B Shingle) is: 0.9172
The shapiro test P-value for sample B (B Shingle) is: 0.0200

(Ref: Python code data file)

From the Shapiro test, as per our assumptions for Sample B the P-value < alpha (Reject the null hypo thesis)

4. Inference Test:

1.Parametric test -Method (Z test, t test)

2.Non parametric test- Method (wilcoxon test)

In our case the data looks normal and the population standard deviation is not know, so we are approaching with parametric test (t test)

Since the sample size are different, we are using one sample t test for Sample A and sample B

The T-test results for Sample A (shingle A) is
The test-statistic is: -1.473
For the one tailed t test the P-value is: 0.0747

From t test one sample the P-value is greater than alpha.(0.07 > 0.05)

The T-test results for Sample B (shingle B) is
The test-statistic is: -3.100
For the one tailed t test the P-value is: 0.0020

From t test one sample the P-value is greater than alpha.(0.07 < 0.05)

5. Decision:

3.1 Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

For SampleA (Shingles-A):

From the Inference test the P-value is greater than alpha (pvalue > α 0.05).
So, we are Fail to Reject the Null Hypothesis Hoo,

Hence for Shingles -A the mean moisture content present in the sample is equal to within permissible limits (0.35 pound per 100 square feet)

For SampleB (Shingles-B):

From the Inference test the P-value is less than alpha (pvalue < α 0.05).
So, we Reject the Null Hypothesis Ho.

Hence for Shingles -B the mean moisture content present in the sample is less than the permissible limits (0.35 pound per 100 square feet)

3.2 Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

1.Hypothesis formulation :

decide null and alternate hypothesis:

Ho: μa - μb = 0
Null Hypothesis Ho states that the mean moisture content present in Shingle A is equal to Shingle B

Hα : μa - μb ≠ 0
Alternate Hypothesis Hα states that the mean moisture content present in Shingle A is not equal to Shingle B

2. Decide the level of significance:

The level of significance is not mentioned in problem statement, so generally we consider 5%
α = 0.05

3. Assumptions:

Test of Means using two methods:

3.1) Histogram (visual Test)

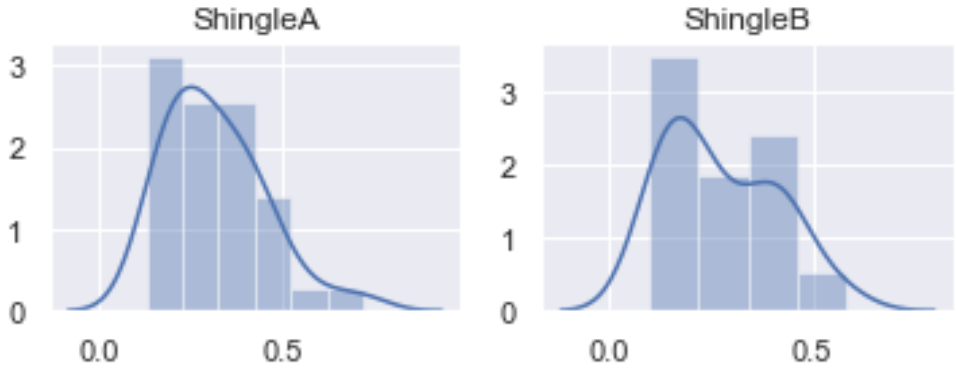
3.2) Shapiro test (Statistical Test)

Test of equality of variance in the data by two methods:

3.4) Box Plot (Visual method)

3.5) Levene Test (Statistic method)

3.1 Histogram:



From the above histogram, The Shingle A - visually the data looks normally distributed.

The Shingle B - visually the data looks like slightly right skewed distribution

3.2 Shapiro Test:

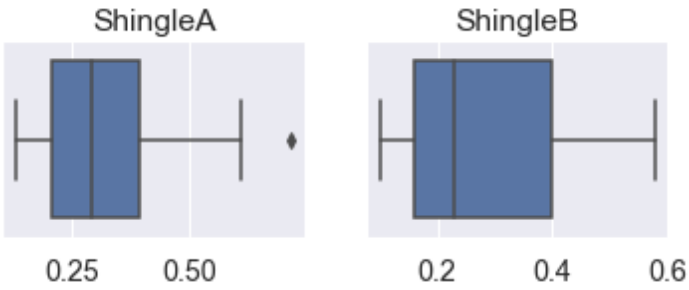
The shapiro test-statistic for sample A (A Shingle) is: 0.9375
The shapiro test P-value for sample A (A Shingle) is: 0.0426
(Ref: Python code data file)

From the Shapiro test, as per our assumptions for Sample A the P-value < alpha.

The shapiro test-statistic for sample B (B Shingle) is: 0.9172
The shapiro test P-value for sample B (B Shingle) is: 0.0200
(Ref: Python code data file)

From the Shapiro test, as per our assumptions for Sample B the P-value < alpha (Reject the null hypo thesis)

3.3 Box Plot:



3.4 Levene Test:

The test-statistic is: 0.238
The P-value is: 0.6272

(Ref: Python code data file)
From the above Levene test states the P-value is greater than α (0.05), hence going for parametric test.

4. Inference Test:

1.Parametric test -Method (Z test ,t test)

2.Non parametric test- Method (mannwhitneyu test)

In our case the variance is looks equal and the population standard deviation is not known, so we are approaching with two sample t test

T-test two sample the following results are obtained:
The test-statistic is: 1.289
The Pvalue is: 0.201

From the t test the P-value is greater than alpha.

5. Decision:

3.2 Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

From the Inference test(t test) the P-value is greater than alpha (pvalue > α 0.05).

So we fail to Reject the Null Hypothesis H*o*,

Hence the population means for Shingles A and B are equal.

The assumptions made to confirm the Inference test is correct and the values from equality of means (shapiro test) and variance in data (levene test) are state to reject the Null hypothesis.