10/07/2022

SMDM Project

BUSINESS REPORT

By,
Balasubramaniyam Ravichandran

Problem 1

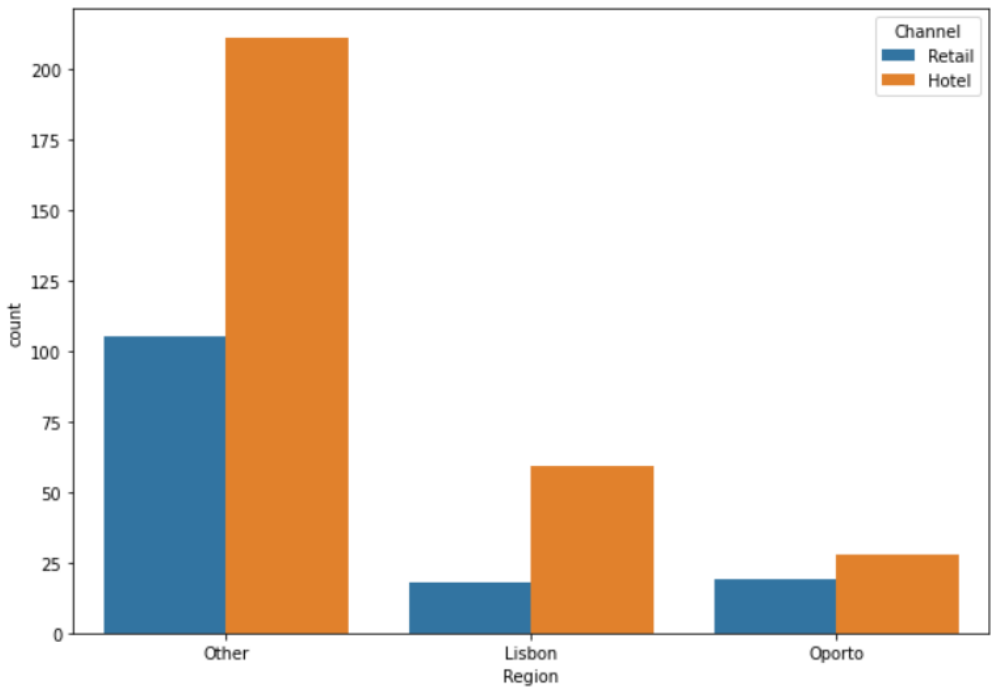
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon,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

1.1 Use methods of descriptive statistics to summarize data.
Which Region and which Channel spent the most?
Which Region and which Channel spent the least?

The available dataset has details about the variety of items sold by our wholesale distributor in different regions and channels. The variety of items sold by our wholesaler are as follows,

- Fresh
- Milk
- Grocery
- Frozen
- Detergent paper
- Delicatessen

Bar graph for spending across different regions and channels is below,



From the Bar graph, we can conclude the following:

i. The most spending is done at other region and the hotel channel.
ii. The least spending is done at Oporto region and the retail channel.

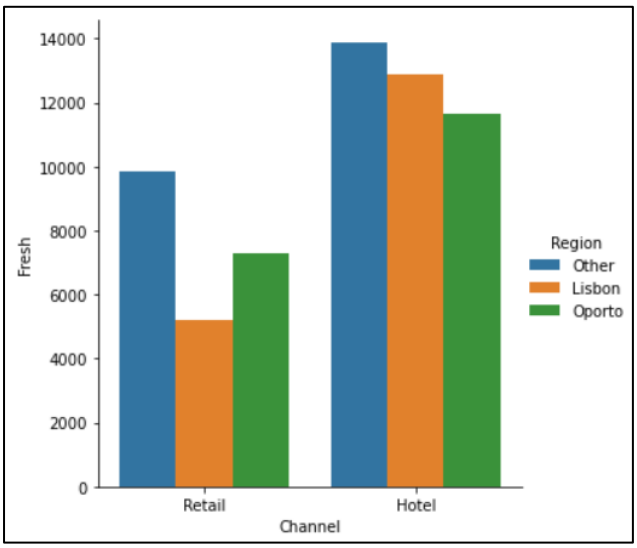
1.2 There are 6 different varieties of items that are considered.
Describe and comment/explain all the varieties across Region and Channel?
Provide a detailed justification for your answer.

The variety of items sold by our wholesaler are as follows,

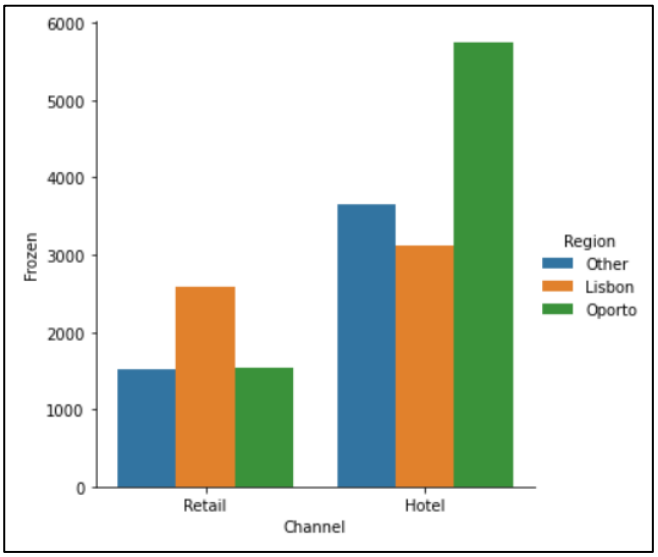
- Fresh
- Milk
- Grocery
- Frozen
- Detergent paper
- Delicatessen

The statistical summary of the above items is as follows in the form of a bar graph,

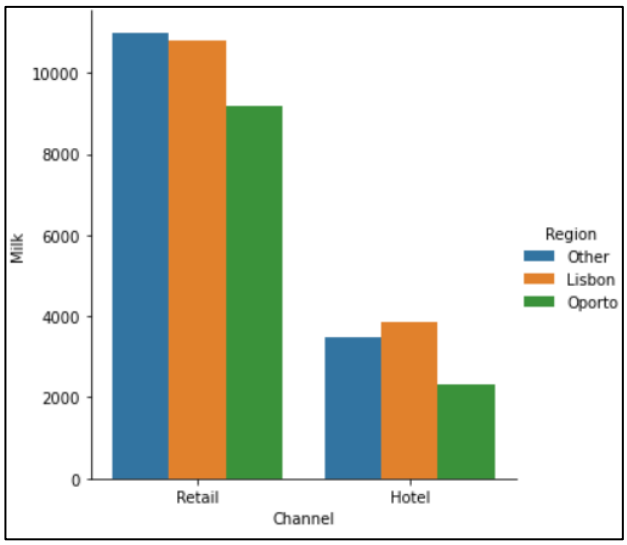
FRESH



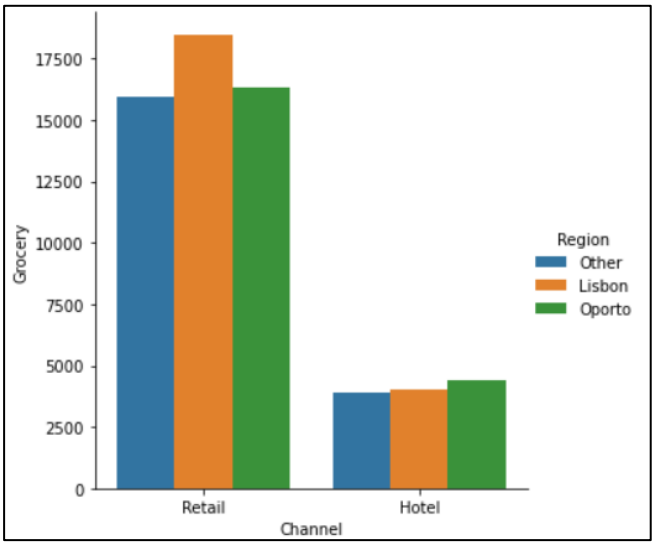
FROZEN



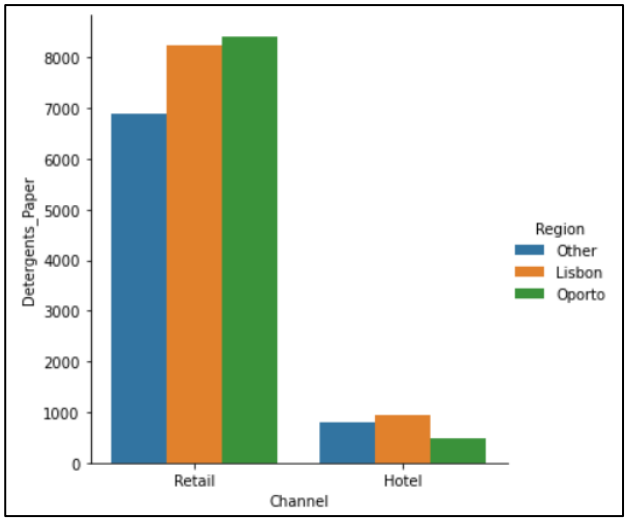
MILK



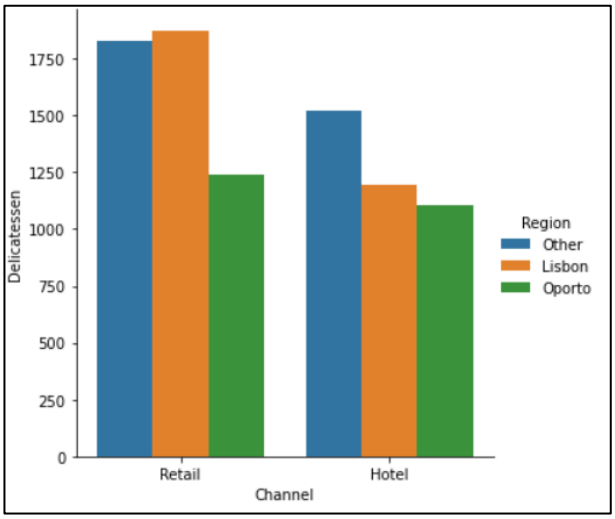
GROCERY



Detergent paper



Delicatessen



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

By measuring the coefficient of variation (CV), we can easily find out the consistency behaviour of our items. The smaller the CV, the higher is the consistency.

Table for CV of all the items,

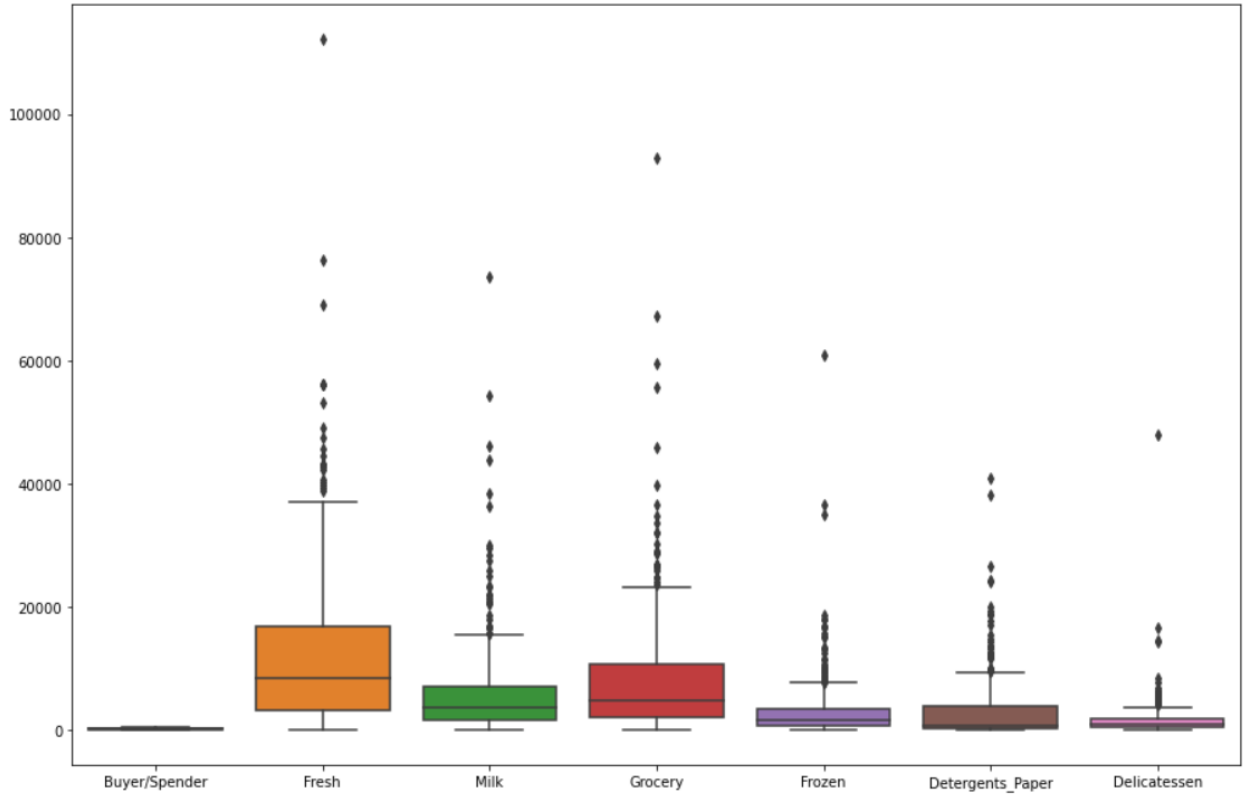
Fresh	Milk	Grocery	Frozen	Detergents Paper	Delicatessen
1.052719608	1.271850831	1.193815448	1.57853553	1.652765788	1.847304104

The Delicatessen item is showing an inconsistent behaviour with (CV of 1.847304).
The Fresh items are showing a consistent behaviour with (CV of 1.052720).

1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

There are outliers present in our data and it is present in all 6 variety of items. These outliers can be identified by using the boxplot technique and the below plot indicates the distribution of data of all 6 variety of items are right skewed.

The box plot for our dataset is as follows,



1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.

Our wholesaler has to try and promote the sales in Oporto region and the retail channel. New promotional offers may be used to attract customers and make them spend more in this region and channel. Delicatessen products sales are to be monitored for its sales during a particular season or event to improve its sale during those periods of times.

Problem 2 -

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.

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

2.1.1. Gender and Major

Major / Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

2.1.2. Gender and Grad Intention

Grad Intention / Gender	No	Undecided	Yes
Female	9	13	11
Male	3	9	17

2.1.3. Gender and Employment

Employment / Gender	Full-Time	Part-Time	Unemployed
Female	3	24	6
Male	7	19	3

2.1.4. Gender and Computer

Computer / Gender	Desktop	Laptop	Tablet
Female	2	29	2
Male	3	26	0

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

Based on the data, answer the following question:

From the data,

The total students = 62

Total male students = 29

Total female students = 33

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

Probability for male student = (Total male students/ The total students)

The probability that a randomly selected CMSU student will be male is 46.78 %

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

Probability for female student = (Total female students/ The total students)

The probability that a randomly selected CMSU student will be Female is 53.23%

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

Based on the data, answer the following question:

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

Table for gender and major,

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Probability of majors for male students = (Male students in a major/ Total male students)

Solution:

The probability of male students in Accounting is: 13.8 %
The probability of male students in CIS is: 3.45 %
The probability of male students in Economics/Finance is: 13.8 %
The probability of male students in International Business is: 6.9 %
The probability of male students in Management is: 20.7 %
The probability of male students in Other is: 13.8 %
The probability of male students in Retailing/Marketing is: 17.25 %
The probability of male students in Undecided is: 10.35 %

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

By referring the table for gender and major,

Probability of majors for female students = (female students in a major/ Total female students)

Solution:

The probability of Female students in Accounting is: 9.1 %
The probability of Female students in CIS is: 9.1 %
The probability of Female students in Economics/Finance is: 21.22 %
The probability of Female students in International Business is: 12.13 %
The probability of Female students in Management is: 12.13 %
The probability of Female students in Other is: 9.1 %
The probability of Female students in Retailing/Marketing is: 27.28 %
The probability of Female students in Undecided is: 0.0 % (all female students have decided their major)

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

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

Table for gender and intention to graduate,

Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

Probability of male students intends to graduate = (Male students yes to graduate/ Total male students) * (Total male students/ The total students)

Solution:

The probability of a randomly chosen male student who intends to graduate is 27.42 %

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

Table for gender and computer,

Computer Gender	Desktop	Laptop	Tablet
Female	2	29	2
Male	3	26	0

Probability of female student without laptop = (female without laptop/ Total female students) * (Total female students/ The total students)

Solution:

The probability of a randomly chosen female student without laptop is 6.45 %

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

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

Table for gender and Employment,

Employment Gender	Full-Time	Part-Time	Unemployed
Female	3	24	6
Male	7	19	3

From the table for gender and Employment,
P(male) = (Total male students/ The total students)

P(male full-time) = (7/ The total students)

P(full-time employed) = (10/ The total students)

Formula= P(male)+ P(full-time employed)- P(male full-time)

Solution:

The probability that a randomly chosen student is a male or has full-time employment: 51.6 %

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

Table for gender and major,

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

From the table,

P(Female in international business) = (4/ Total female students)

P(Female in management) = (4/ Total female students)

Formula = P(Female in international business) + P(Female in management)

Solution:

The probability a female student is major in International Business or Management is 24.25 %

2.6. Construct a contingency table of Gender & 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?

Table for gender and intention to graduate (Yes/No),

Grad Intention Gender	No	Yes
Female	9	11
Male	3	17

From the table,

p(grad yes) = ((11+17)/ The total students)
p(female) = (Total female students / The total students)
Probability of being female and intending to graduate (AnB)
P(AnB) = ((11/ Total female students) * (Total female students / The total students))

Solution:

P(AnB)≠ p(grad yes)* p(female)

We can conclude that the graduate intention and being female are not independent 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?

Formula:
P(GPA<3) = (Total students with GPA<3) / (The total students)

Solution:
Total students with (GPA<3) = 17
The total students = 62
The probability that a student's GPA is less than 3 is 27.42 %

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.

Table for gender and salary≥50,

Gender	Female	Male
Salary		
False	15	15
True	18	14

Formula:

From the table for gender and salary≥50
For male,
P(S≥50) = (14/ Total male students)
For female,
P(S≥50) = (18/ Total female students)

Solution:
The probability that a randomly selected male earns 50 or more is 48.28%
The probability that a randomly selected female earns 50 or more is 54.55 %

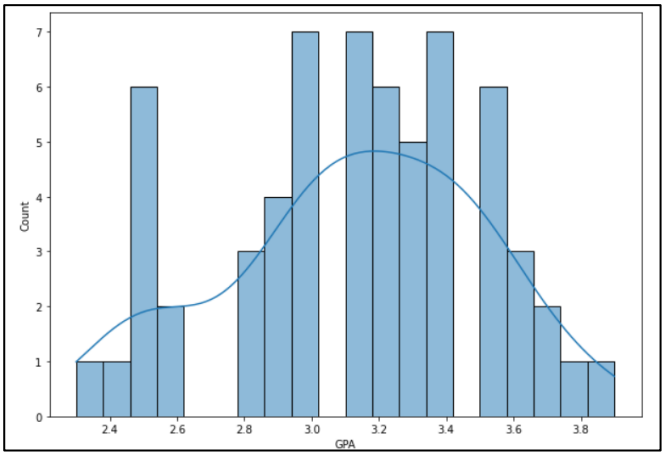
2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.

We can find out if our dataset for GPA, Salary, Spending, and Text Messages are normally distributed or not by plotting a histogram and observing the histogram's KDE. For a normally distributed data the 'Bell Shaped Distribution' is to be-

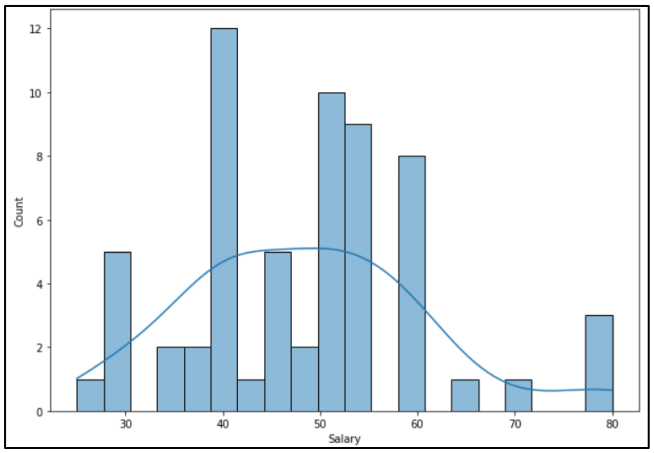
-observed, it is symmetrical about its mean and the mean, the median, and the mode are all equal.

Histograms for GPA, Salary, Spending, and Text Messages are below,

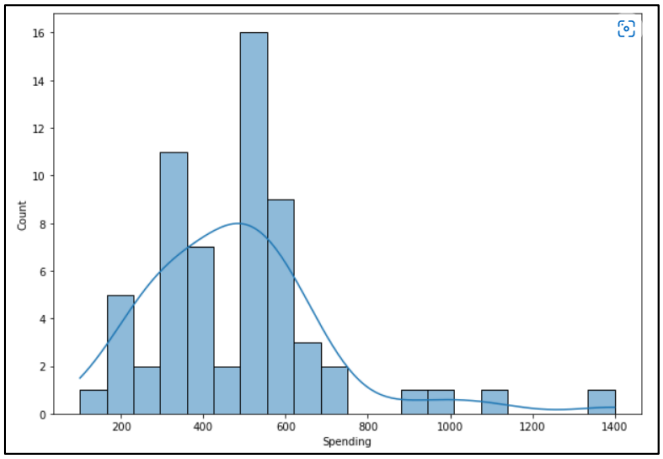
GPA



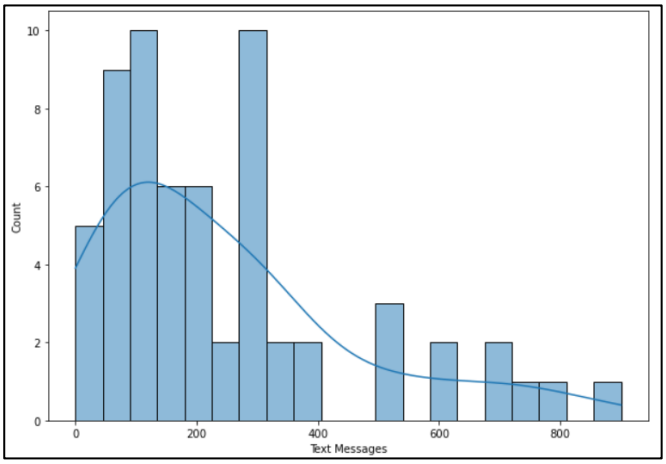
SALARY



SPENDING



TEXT MESSAGES



Conclusion for normal distribution:

1. The distribution of GPA is not normally distributed in our dataset and has multiple modes. The mean is 3.129032, median is 3.15 and has multiple modes (3, 3.1, 3.4).

2. The distribution of Salary is not normally distributed in our dataset. It is right skewed, the mean is 48.548387, median is 50.00 and mode is 40

3. The distribution of Spending is not normally distributed in our dataset. It is right skewed, the mean is 482.016129, median is 500.00 and mode is 500

4. The distribution of Text Messages is not normally distributed in our dataset. It is right skewed, the mean is 246.209677, median is 200.00 and mode is 300

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 coloring 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 pounds per 100 square feet.

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

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

SOLUTION:

Step 1: Define null and alternative hypotheses for our problem, same for both

Shingles (A and B).

Null hypothesis, H0: $\mu <= 0.35$

Alternative hypothesis, H1: $\mu > 0.35$

Step 2: Decide the significance level

Here we select $\alpha = 0.05$.

Step 3: Identify the test statistic

We need to perform t-test(one-sample) for both Shingle-A and Shingle-B

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

$\mu = 0.35$

X-bar = sample mean = (0.316- for Shingle-A, 0.27- for Shingle-B)

S = sample std.dev = 0.136 for Shingle-A, 0.14- for Shingle-B)

n = sample size = 36 for Shingle-A, 31 for Shingle-B)

Step 4: Calculate the p - value and test statistic

By using the formula mentioned in identify the test statistics, we are calculating the p-value.

For Shingle-A,
One sample t test
t statistic: -1.4735046253382782 p value: 0.07477633144907513
Level of significance: 0.05
We have no evidence to reject the null hypothesis for Shingle_A since p value > Level of Significance.

For Shingle-B,
One sample t test
t statistic: -3.1003313069986995 p value: 0.0020904774003191813
Level of significance: 0.05
We have evidence to reject the null hypothesis for Shingle_B since p value < Level of Significance.

Step 5: Conclusion

For Shingle-A,

the mean moisture content is less than 0.35 pounds per 100 square feet.

For Shingle-B,

the mean moisture content is not less than 0.35 pounds 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?

Step 1: Define null and alternative hypotheses for our problem,¶

Null hypothesis, H0: $\mu(A) = \mu(B)$

Alternative hypothesis, H1: $\mu(A) \neq \mu(B)$

Step 2: Decide the significance level

Here we select $\alpha = 0.05$.

Step 3: Identify the test statistic

We need to perform t-test(two-sample) for both Shingle-A and Shingle-B

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

X-bar1, X-bar2 = sample mean = (0.316- for Shingle-A, 0.27- for Shingle-B)

S1, S2 = sample std.dev = 0.136 for Shingle-A, 0.14- for Shingle-B)

n1, n2 = sample size = 36 for Shingle-A, 31 for Shingle-B)

Step 4: Calculate the p - value and test statistic

By using the formula mentioned in identify the test statistics, we are calculating the p-value.
P Value 0.00209(from python)
Level of significance: 0.05
We have evidence to reject the null hypothesis, since p value < Level of Significance.

Step 5: Conclusion

The population means for shingles A and B are not equal.

Assumptions,

1. The null and alternative hypothesis are to be framed
2. Level of significance: 0.05