SMDM PROJECT

REPORT

DSBA

Contents

[Problem 1 2](#Problem1)

[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? 2](#Problem1a)

[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 2](#Problem12)

[1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour? 2](#Problem13)

[1.4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments 2](#Problem14)

[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 3](#Problem15)

[Problem 2 3](#_Toc6894)

[2.1. For this data, construct the following contingency tables (Keep Gender as row variable) 3](#_Toc6895)

[2.1.2. Gender and Grad Intention 3](#_Toc6896)

[2.1.3. Gender and Employment 3](#_Toc6897)

[2.1.4. Gender and Computer 3](#_Toc6898)

[2.2. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following questions: 3](#_Toc6899)

[2.2.1. What is the probability that a randomly selected CMSU student will be male? What is the probability that a randomly selected CMSU student will be female? 3](#_Toc6900)

[2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender? Justify your comment in each case 4](#_Toc6901)

[2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions. [Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric] 4](#_Toc6902)

[Problem 3 5](#_Toc6903)

[3.1 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? 5](#_Toc6904)

[3.2 What assumption about the population distribution is needed in order to conduct the hypothesis tests above? 5](#_Toc6905)

# **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 (Wholesale Customer.csv) 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).

# **Below is the Sample of the dataset that we have used for performing analysis:**

Table

Description automatically generated

# **Below is some information about the columns of the dataset:**

0 Buyer/Spender 440 non-null int64

1 Channel 440 non-null object

2 Region 440 non-null object

3 Fresh 440 non-null int64

4 Milk 440 non-null int64

5 Grocery 440 non-null int64

6 Frozen 440 non-null int64

7 Detergents\_Paper 440 non-null int64

8 Delicatessen 440 non-null int64

From the above Table we can say that all variables are Quantitative except Channel and Region which are Qualitative.

# 

# 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?

# **DESCRIPTIVE STATISTICS TO SUMMARIZE DATA**

Table

Description automatically generated

* A Data Table organizes the data into Individuals and Variables. Following are the characteristics of the ‘Wholesale Cutomer.csv’ Dataset:

Total number of Variables in the data set --> 9

The variables are as follows --> ['Buyer/Spender', 'Channel', 'Region', 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents\_Paper', 'Delicatessen']

Total number of Categorical Variables in the data set --> 2

The Categorical Variables are as follows --> ['Region', 'Channel']

* + The methods of Descriptive Statistics that can help us to Summarize data in a data set are – Mean, Median and Mode. These are the different ways that we can use to describe the ‘Middle’, ‘Center’ or the ‘Typical’ value of the data.
  + Following is the Mean, Median and Mode for all variables:

A) Fresh --->

Mean = 12000.297727272728

Median = 8504.0

On calculating Mode for column 'Fresh', we find that the column has more than one Mode i.e.: Mode= [3, 514, 3366, 7149, 8040, 9670, 18044].

Therefore, mode can't be used a measure of central tendency

B) Milk --->

1. Mean = 5796.265909090909
2. \Median = 3627.0
3. On calculating Mode for column 'Milk', we find that the column has more than one Mode i.e.: Mode= [577, 659, 829, 899, 944, 1012, 1032, 1115, 1196, 1610, 1897, 2428, 2884, 3045, 3199, 3587, 3880, 4230, 5139].
4. Therefore, mode can't be used a measure of central tendency

C) Grocery --->

Mean = 7951.277272727273

Median = 4755.5

On calculating Mode for column 'Grocery', we find that the column has more than one Mode i.e.: Mode= [683, 1493, 1563, 1664, 2062, 2405, 2406, 3600, 6536, 10391].

Therefore, mode can't be used a measure of central tendency

D) Frozen --->

Mean = 3071.931818181818

Median = 1526.0

On calculating Mode for column 'Frozen', we find that the column has more than one Mode i.e.: Mode= [ 133, 364, 402, 425, 744, 779, 824, 830, 848, 937, 1285, 1619, 2540, 4324].

Therefore, mode can't be used a measure of central tendency

E) Detergents\_Paper --->

Mean = 2881.4931818181817

Median = 816.5

On calculating Mode for column 'Detergents\_Paper', we find that the column has more than one Mode i.e.: Mode= [3, 20, 56, 69, 70, 93, 96, 118, 153, 182, 210, 212, 227, 232, 256, 284, 311, 397, 483, 788, 811, 918, 955].

.Therefore, mode can't be used a measure of central tendency

F) Delicatessen --->

* + - 1. Mean = 1524.8704545454545,
      2. Median = 965.5
      3. On calculating Mode for column 'Delicatessen', we find that the column has two Modes i.e.: Mode= [3, 834].
      4. Therefore, mode can be used a measure of central tendency.
      5. The column 'Delicatessen' is Bimodal

# **REGION wise total spend**

Chart, bar chart

Description automatically generated

As is evident from the above Figure we can conclude as follows:

* + - Region - OTHER spends the most
    - Region - OPORTO spends the least

# **CHANNEL wise total spend**

Chart, bar chart

Description automatically generated

As is evident from the above Figure we can conclude as follows:

1. Channel - HOTEL spends the most
2. Channel - RETAIL spends the least

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

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

* The above figures shows mean spends of all items across Regions and Channels.
* The items Frozen, Milk, Grocery, Delicatessen and Fresh vary across Channels.
* The variation for item Detergents\_Paper across Channels is highest.
* For items Frozen and Fresh, Hotels spends the most. For items Milk, Grocery, Delicatessen and Detergent\_Paper Retail spends the most.

# **Following is the analysis of each item across Region and Channel:**

* + - FROZEN
      * Hotel has spent more on item FROZEN as compared to Retail.
      * There is not much variation in the spent amount for item FROZEN across Channels.
    - MILK
      * Retail has spent more on item MILK as compared to Hotel.
      * There is a higher variation in the spent amount for item MILK across Channels.
    - GROCERY
      * Retail has spent more on item GROCERY as compared to Hotel.
      * There is a higher variation in the spent amount for item GROCERY across Channels.
    - DELICATESSEN
      * Retail has spent more on item DELICATESSEN as compared to Hotel.
      * There is not much variation in the spent amount for item DELICATESSEN across Channels.
    - FRESH
      * Hotel has spent more on item FRESH as compared to Retail.
      * There is some variation in the spent amount for item FRESH across Channels.
    - DETERGENT\_PAPER
      * Retail has spent more on item DETERGENT\_PAPER as compared to Hotel.
      * There is very high variation in the spent amount for item DETERGENT\_PAPER across Channels.

**Region-Wise Analysis**

* + Lisbon

Table

Description automatically generated

* The count of Spent is 77.
* The Standard deviation varies from 1345.4233 to 11557.438
* The minimum spent is 5.00 for item Detergents\_Paper
* The maximum spent is 56083.00 for item Fresh
  + Oporto

Table

Description automatically generated

* The count of Spent is 47.
* The Standard deviation varies from 1050.739 to 10842.745
* The minimum spent is 3.00 for item Fresh
* The maximum spent is 67298.00 for item Grocery
  + Other

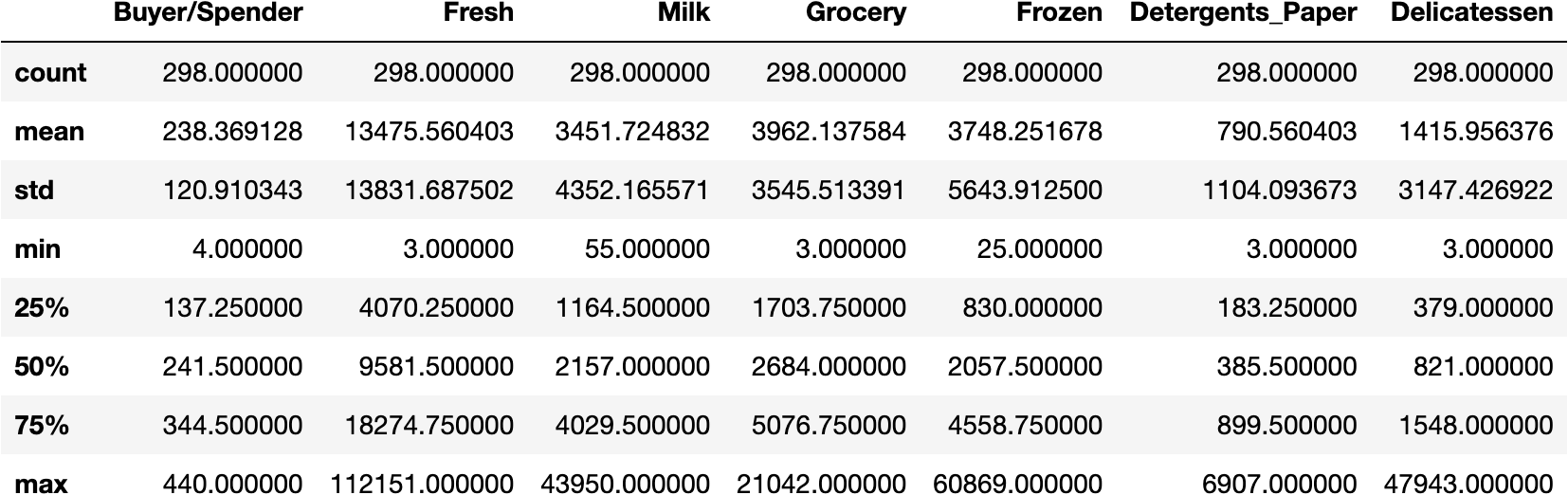
Table

Description automatically generated

* The count of Spent is 316.
* The Standard deviation varies from 3232.581 to 13389.213
* The minimum spent is 3.00 for item Fresh, Grocery, Detergents\_Paper and Delicatessen
* The maximum spent is 112151.00 for item Fresh

**Channel-Wise Analysis**

a) Hotel



* The count of Spent is 298.
* The Standard deviation varies from 1104.0936 to 13831.6875
* The minimum spent is 3.00 for item Fresh, Grocery, Detergents\_Paper and Delicatessen
* The maximum spent is 112151.00 for item Fresh
  1. Retail

Table

Description automatically generated

* The count of Spent is 142.
* The Standard deviation varies from 1812.80 to 12267.3180
* The minimum spent is 3.00 for item Delicatessen
* The maximum spent is 92780.00 for item Grocery

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?

* Descriptive Measure of Variability is the Spread of the data and can be calculated using standard deviation, variance, minimum and maximum variables.
* Standard Deviation of each item was calculated
  + Standard Deviation of item Fresh = 12647.328865076894
  + Standard Deviation of item Grocery = 9503.162828994346
  + Standard Deviation of item Milk = 7380.377174570843
  + Standard Deviation of item Frozen = 4854.673332592367
  + Standard Deviation of item Detergent\_Paper = 4767.8544479042
  + Standard Deviation of item Delicatessen = 2820.1059373693975
* We used the Standard Deviation values to calculate the Coefficient of Variation for each item
  + Coefficient of Variation of FRESH = 1.0539179237473149
  + Coefficient of Variation of GROCERY = 1.1951743730016824
  + Coefficient of Variation of MILK = 1.2732985840065414
  + Coefficient of Variation of FROZEN = 1.5803323836352914
  + Coefficient of Variation of DETERGENTS\_PAPER = 1.6546471385005155
  + Coefficient of Variation of DELICATESSEN = 1.8494068981158382

# **We observed the following from the above analysis:**

* Maximum Coefficient of Variation is 1. 8494068981158382 for item DELICATESSEN. This means DELICATESSEN item shows the most inconsistence behaviour.
* Minimum Coefficient of Variation is 1.0539179237473149 for item FRESH. This means FRESH item shows the least inconsistence behaviour.

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

Chart, box and whisker chart

Description automatically generated

# **From the above figure we can conclude the following:**

* There are Outliers observed in all six items.
* All the Outliers are greater than the maximum value and therefore their mean is greater than the median.
* The distribution is Right-Skewed.

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

Table

Description automatically generated

* From the above Cross Table, we can observe that the Region – Other has more than 70% of the Total spenders. Around 10% of the Spenders lie in the Region Oporto and around 17% of the Spenders are in Region Lisbon. This also means that we have no knowledge regarding the Region of 70% of the Spenders in our Data set.
* Around 67% of the Total Spenders are spending via hotel and around 32% of Total Spenders are spending via Retail.
* The Business should target increasing sales in the Region Others as they constitute 70%of the total spenders. The Business should also start promoting their items in the Region Oporto and Lisbon since there are only 30% spenders. The aim should be to attract more and more spenders so that the sale can increase in these regions as well.

Chart

Description automatically generated

* From the above Heatmap, we can conclude that the Strongest correlation is observed between Detergents\_Paper and Grocery, Milk and Grocery, Milk and Detergents\_Paper. The Weakest correlation is observed between Frozen and Detergents\_Paper, Fresh and Grocery, Frozen and Grocery.
* The items for which we have positive correlation or strong correlation, the business should start considering creating offers which would include these items in combination. This would result in increase in profits by increasing the sale.

# **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 (stored in the Survey data set).

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

**2.1.1. Gender and Major**

Table

Description automatically generated

**2.1.2. Gender and Grad Intention**

Table

Description automatically generated

**2.1.3. Gender and Employment**

Table

Description automatically generated

**2.1.4. Gender and Computer**

Table

Description automatically generated

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

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

Total number of students = 62

Total number of male students = 29

Probability that a randomly selected CMSU student will be male = 29/62 i.e., 0.4677

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

Total number of students = 62

Total number of female students = 33

Probability that a randomly selected CMSU student will be female = 33/62 i.e., 0.5322

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

Description automatically generated

From the above Crosstab, we can observe the following:

Total number of male students across different majors = 29

Total number of male students in accounting = 4

Probability of accounting major among male = 4/29 i.e., 0.138

Total number male students in CIS = 1

Probability of CIS major among male = 1/29 i.e., 0.034

Total number of male students in Economics/Finance = 4

Probability of Economics/Finance major among male = 4/29 i.e., 0.138

Total number of male students in International Business = 2

Probability of International Business major among male = 2/29 i.e., 0.0690

Total number of male students in Management = 6

Probability of Management major among male = 6/29 i.e., 0.207

Total number of male students in Other = 4

Probability of Other major among male = 4/29 i.e., 0.138

Total number of male students in Retailing/Market = 5

Probability of Retailing/Market major among male = 5/29 i.e., 0.172

Total number of male students in Undecided = 3

Probability of Undecided major among male = 3/29 I.e., 0.103

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

Table

Description automatically generated

From the above Crosstab, we can observe the following:

Total number of female students across different majors = 33

Total number of female students in accounting = 3

Probability of accounting major among male = 3/33 i.e., 0.09

Total number female students in CIS = 3

Probability of CIS major among female = 3/33 i.e., 0.09

Total number of female students in Economics/Finance = 7

Probability of Economics/Finance major among female = 7/33 i.e., 0.212

Total number of female students in International Business = 4

Probability of International Business major among female = 4/33 i.e., 0.121

Total number of female students in Management = 4

Probability of Management major among female = 4/33 i.e., 0.121

Total number of female students in Other = 3

Probability of Other major among female = 3/33 i.e., 0.09

Total number of female students in Retailing/Market = 9

Probability of Retailing/Market major among female = 9/33 i.e., 0.0.272

Total number of female students in Undecided = 0

Probability of Undecided major among female = 0/33 I.e., 0.0

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.

P(A and B) = P(A) \* P(B)

where,

P(A) = Probability that a randomly selected student is a male student

P(B) = Probability that a randomly selected student intends to graduate

P(A and B) = Probability that a randomly selected student is a male student and intends to graduate

Table

Description automatically generated

From the above table we can observe the following:

P(A) = 29/62 i.e., 0.467

P(B) = 17/29 i.e., 0.586

P(A and B) = 0.467 \* 0.586 = 0.2736

Therefor the Probability that a randomly chosen is a male and intends to graduate is 0.2736

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

P(A and B) = P(A) \* P(B)

where,

P(A) = Probability that a randomly selected student is a female student

P(B) = Probability that a randomly selected student who does not have a laptop

P(A and B) = Probability that a randomly selected student is a female student and does not have a laptop

Table

Description automatically generated

From the above table we can observe the following:

P(A) = 33/62 i.e., 0.5322

P(B) = 4/33 i.e., 0.121

P(A and B) = 0.5322 \* 0.121 = 0.0643

Therefore the Probability that a randomly selected student is a female and does not have a laptop is 0.0643

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?

P(A or B) = P(A) + P(B) – P(A and B)

where,

P(A) = Probability that a randomly selected student is a male

P(B) = Probability that a randomly selected student has a full employment

P(A and B) = Probability that a randomly selected student is a male and has a full-time employment

P(A or B) = Probability that a randomly selected student is a male or has a full-time employment

Table

Description automatically generated

From the above table we can observe the following:

P(A) = 29/62 i.e., 0.467

P(B) = 10/62 i.e., 0.161

P(A and B) = 7/62 i.e., 0.112

P(A or B) = 0.467 + 0.161 – 0.112 = 0.516

Therefore the Probability that a randomly chosen student is a male or has a full-time employment is 0.516

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

P(A or B) = P(A) + P(B) – P(A and B)

where,

P(A) = Probability that a randomly selected female student is majoring in International Business

P(B) = Probability that a randomly selected female student is majoring in management

P(A and B) = Probability that a randomly selected female student is majoring in International Business and Management

P(A or B) = Probability that a randomly selected female student is majoring in International Business or Management

Table

Description automatically generated

From the above table we can observe the following:

P(A) = 4/33 i.e., 0.121

P(B) = 4/33 i.e., 0.121

P(A and B) = 0, since it is not possible to do more than one major at a time.

P(A or B) = 0.121 + 0.121 – 0 = 0.242

Therefore the probability that given a female student is randomly chosen, she is majoring in International Business and Management is 0.242.

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?

**Table

Description automatically generated**

The table shown above is a contingency table of Gender and Intent to Graduate at 2 levels(Yes/No). The undecided students are not considered In the above table.

If P(A and B) = P(A) \* P(B) then A and B can be called Independent events.

Here P(A) = Probability that selected student intends to graduate = 28/40

P(B) = Probability that selected student is a female = 20/40

P(A) \* P(B) = 28/40 \*20/40 = 0.35

From the table we can say that,

P(A and B) = 11/20 \* 20/40 = 0.275

Since P(A and B) is not equal to P(A) \* P(B), A and B are not Independent Events

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

Answer the following questions based on the data

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

Total number of Students = 62

Total number of Students whose GPA is less than 3 = 17

Probability that his/her GPA is less than 3 = (Total number of Students)/(Total number of Students with GPA<3)

= 17/62 = 0.274

Therefore, if randomly chosen, the absolute probability that his/her GPA is less than 3 is 0.274.

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

Description automatically generated

From the above cross table we can observe the following:

Total number of male that earn more than 50.0 = 14

Total number of male = 15 + 14 = 29

Probability that a randomly selected male earns 50 or more = 14/29 = 0.483

Therefore the conditional probability that a randomly selected male earns 50 or more is 0.483.

Total number of female that earn more than 50.0 = 18

Total number of female = 15 + 18 = 33

Probability that a randomly selected female earns 50 or more = 18/33 = 0.545

Therefore the conditional probability that a randomly selected female earns 50 or more is 0.4.

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.

* GPA

Chart, histogram

Description automatically generated

* + - 1. Mean - 3.129032
      2. Median - 3.150000
      3. Mode - 3.0
      4. Skewness - -0.31460

Therefore we can say that the variable GPA nearly follows Normal Distribution but is

slightly skewed to the left. Also, mean is less than the median.

* + Salary

Chart, histogram

Description automatically generated

* 1. Mean - 48.548387
  2. Median - 50.000000
  3. Mode - 40.0
  4. Skewness - 0.53470

Therefore we can say that the variable Salary does not follow Normally Distribution

But is slightly skewed to the Right. Also, mean is slightly less than the median.

* + Spending

Chart, histogram

Description automatically generated

1. Mean - 482.016129
2. Median - 500.000000
3. Mode - 500
4. Skewness - 1.5859147

Therefore we can say that the variable Spending does not follow Normal Distribution. But is skewed to the right. Also, mean is less than the median.

* + Text Messages

Chart, histogram

Description automatically generated

1) Mean - 246.209677

* + 1. Median - 200.000000
    2. Mode - 300
    3. Skewness - 1.295807

Therefore we can say that the variable Text Messages does not follow Normal Distribution But is slightly to the right. Also, mean is greater than the Median.

According to the Empirical Rule, to achieve Normal Distribution nearly all the data should fall within three Standard Deviations from the mean.

* 68% of the data should fall within the first Standard Deviation from the Mean.
* 95% of the data should fall within the Second Standard Deviation from the Mean.
* 99.7% of the data should fall within the Third Standard Deviation from the Mean.

After applying the above rule to our data set , we found that for variable Salary, Spending and Text Messages do not follow 68-95-99.7 Rule.

# **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 pounds per 100 square feet.

The file ([A & B shingles.csv](https://olympus.mygreatlearning.com/courses/78045/files/5473565/download?verifier=VbWoM7RKLoN2SSc4ab7B8KN97s230sqOqgDf3T3g&wrap=1)) 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 permissiblelimits? State your conclusions clearly showing all steps.

**Shingles A**

To check whether there is evidence that means moisture contents in Shingle A are within the permissible limits, following are the Null and Alternate Hypothesis:

Ho *μ* <= 0.35

Ha *μ*  > 0.35

**Level of Significance**

As the level of significance is not mentioned we can assume as follows:

𝛼 = 0.05

**Test**

* since we need to examine whether the mean of a population is statistically different from a known or hypothesized value, we will be using One Sample T- Test
* We will be using One Tailed Test for Upper Tail

After performing the calculations, the results are as follows:

* Test\_Statistic : -1.473504
* p\_value : 0.07477633144907513

Since p-value is 0.07477633144907513 which is greater than the level of significance, **We Fail to Reject the Null Hypothesis** i.e. for Shingle A the mean moisture content is within the permissible limits.

**Shingles B**

To check whether there is evidence that means moisture contents in Shingle B are within the permissible limits, following are the Null and Alternate Hypothesis:

Ho *μ* <= 0.35

Ha *μ* > 0.35

**Level of Significance**

As the level of significance is not mentioned we can assume as follows:

𝛼 = 0.05

**Test**

* since we need to examine whether the mean of a population is statistically different from a known or hypothesized value, we will be using One Sample T- Test
* We will be using One Tailed Test for Upper Tail

**Result**

* After performing the calculations, the results are as follows:
* Test\_Statistic : -3.1003313069986995
* p\_value : 0.0020904774003191813

**Since p-value is 0.0020904774003191813 which is less than the level of significance, We Can Reject the Null Hypothesis.**

**3.2 Do you think that the population mean 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?**

Following are the Null and Alternate Hypothesis:

Ho : *μa = μb*

Ha : *μa ≠ μb*

**Level of Significance**

As the level of significance is not mentioned we can assume as follows:

𝛼 = 0.05

**Test**

**Assumptions:**

* The data should be Normally Distributed. Central Limit Theorem states that sample mean follows a normal distribution even if population distribution is not normal given the condition that n>=30.
* The sample from the population must be randomly collected.
* For Independent T-test, data of both samples should not have any relation with each other.
* The variables should be continuous.

**Result**

* After performing the calculations, the results are as follows:
  + Test\_Statistic : 1.2896282719661123
  + p\_value : 0.2017496571835306

**Since p-value is greater than level of significance, We do not have enough evidence to reject the Null Hypothesis. We can conclude that the mean Moisture Content in the two sample are same.**