...

SMDM Project Report

- PGPDSBA Online Sep_A 2021
- Student: **Hemant Patidar**

Contents

List of Figures/Tables:

Problem 1 (Wholesale Customers Analysis)

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).

Basic Data Definition:

Buyer/Spender: Serial numbers to identify Buyer or Spender

Channel: Sales channel (either Retail or Hotel)

Region: Sales Regions (either Lisbon, Oporto or Other)

Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen: Annual spending with respect to each variety.

*Assumptions/Corrections: The dataset has no missing values, since we are not assuming or removing any record… And Region Lisbon, Oporto are in Portugal, so we can assume the spending would be in Euros (€).

Total Spending: Total Annual spending of all varieties

1.1 Use methods of descriptive statistics to summarize data

Region	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total Spending
Lisbon	Hotel	761233	228342	237542	184512	56081	70632	1538342
	Retail	93600	194112	332495	46514	148055	33695	848471
Oporto	Hotel	326215	64519	123074	160861	13516	30965	719150
	Retail	138506	174625	310200	29271	159795	23541	835938
Other	Hotel	2928269	735753	820101	771606	165990	320358	5742077
	Retail	1032308	1153006	1675150	158886	724420	191752	4935522

Figure A - Descriptive stats of Wholesale Data

Channel	Total Spending
Hotel	7999569
Retail	6619931

Region	Total Spending
Lisbon	2386813
Oporto	1555088
Other	10677599

Figure B - Region & Channel wise spending

1.1.1 Which Region and which Channel spent the most?

It is evident from above table (Figure A), **Hotels** in **Other** region has most annual spending.

And individually (Figure B), we have most spending in **Other region (€ 10.6 M)** and **Hotel Channel (€ 7.9 M)**

1.1.2 Which Region and which Channel spent the least?

By Figure A, it is noticeable that **Retail** channel in **Lisbon** region has least annual spending.

And individually (Figure B), we have least spending in **Oporto region (€ 1.5 M)** and **Retail Channel (€ 6.6 M)**

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.

Overall summary:

	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
Total Spending	440.0	33226.136364	26356.301730	904.0	17448.75	27492.0	41307.50	199891.0

Figure C - Statistics of different varieties

Region-wise Breakdown:

	Region	Lisbon	Oporto	Other
Fresh	count	77.000000	47.000000	316.000000
	mean	11101.727273	9887.680851	12533.471519
	std	11557.438575	8387.899211	13389.213115
Milk	mean	5486.415584	5088.170213	5977.085443
	std	5704.856079	5826.343145	7935.463443
Grocery	mean	7403.077922	9218.595745	7896.363924
	std	8496.287728	10842.745314	9537.287778
Frozen	mean	3000.337662	4045.361702	2944.594937
	std	3092.143894	9151.784954	4260.126243
Detergents_Paper	mean	2651.116883	3687.468085	2817.753165
	std	4208.462708	6514.717668	4593.051613
Delicatessen	mean	1354.896104	1159.702128	1620.601266
	std	1345.423340	1050.739841	3232.581660

Figure D - Region-wise Stats breakdown of Varieties

Channel-wise Breakdown:

	Channel	Hotel	Retail
Fresh	count	298.000000	142.000000
	mean	13475.560403	8904.323944
	std	13831.687502	8987.714750
Milk	mean	3451.724832	10716.500000
	std	4352.165571	9679.631351
Grocery	mean	3962.137584	16322.852113
	std	3545.513391	12267.318094
Frozen	mean	3748.251678	1652.612676
	std	5643.912500	1812.803662
Detergents_Paper	mean	790.560403	7269.507042
	std	1104.093673	6291.089697
Delicatessen	mean	1415.956376	1753.436620
	std	3147.426922	1953.797047

Figure E - Channel-wise breakdown of Varieties

By above description stats diagrams, we can infer –

We have more Hotels as buyer/spender compared to Retails, and number of stores/(buyer/spender) is more in “Other” region than Oporto or Lisbon.

- **Fresh** has an average spending of € 1,200.29 with standard deviation 12,647.32, medians of 75% (Q3) is € 16,933.75, 50% (Q2) is € 8,504.0 and 25% (Q1) is € 3,127.75… Minimum spend € 3.0, and maximum spend € 112,151.0, and IQR (Q3- Q1) 13,806.0…
    - Average spending in Oporto is lesser compared to Lisbon or Other, and average spend is highest in "Other" region, also spread (std) is highest in "Other" region.
    - Spends varies most in Hotels, and average spending is also high there.

- **Milk** has an average spending of € 5,796.26 with standard deviation 7,380.37, medians of 75% (Q3) is € 7190.25, 50% (Q2) is € 3627.0 and 25% (Q1) is € 1533.0 … Minimum spend € 55.0, and maximum spend € 73,498.0, and IQR (Q3- Q1) 5,657.25…
    - Average spending in Oporto is lesser compared to Lisbon or Other, and average spend is highest in "Other" region, also spread (std) is highest in "Other" region.
    - Spends varies most in Retails, and average spending is also high there.

- **Grocery** has an average spending of € 7,951.27 with standard deviation 9,503.16, medians of 75% (Q3) is € 10,655.75, 50% (Q2) is € 4,755.5 and 25% (Q1) is € 2,153.0 … Minimum spend € 3.0, and maximum spend € 92,780.0, and IQR (Q3- Q1) 8,502.75…
    - Average spending in Lisbon is lesser compared to Oporto or Other, and average spend is highest in "Oporto" region, also spread (std) is highest in "Oporto" region.
    - Spends varies most in Retails, and average spending is also high there.

- **Frozen** has an average spending of € 3,071.93 with standard deviation 4,854.67, medians of 75% (Q3) is € 3,554.25, 50% (Q2) is € 1,526.0 and 25% (Q1) is € 742.25 … Minimum spend € 25.0, and maximum spend € 60,869.0, and IQR (Q3- Q1) 2,812.0…
    - Average spending in "Other" is lesser compared to Oporto or Lisbon, and average spend is highest in "Oporto" region, also spread (std) is highest in "Oporto" region.
    - Spends varies most in Hotels, and average spending is also high there.

- **Detergents Paper** has an average spending of € 2,881.49 with standard deviation 4,767.85, medians of 75% (Q3) is € 3,922.0, 50% (Q2) is € 816.5 and 25% (Q1) is € 256.75 … Minimum spend € 3.0, and maximum spend € 40,827.0, and IQR (Q3- Q1) 3,665.25…
    - Average spending in "Lisbon" is lesser compared to Oporto or Other, and average spend is highest in "Oporto" region, also spread (std) is highest in "Oporto" region.
    - Spends varies most in Retails, and average spending is also high there.

- **Delicatessen** has an average spending of € 1,524.87 with standard deviation 2,820.10, medians of 75% (Q3) is € 1,820.25, 50% (Q2) is € 965.5 and 25% (Q1) is € 408.25 … Minimum spend € 3.0, and maximum spend € 47,943.0, and IQR (Q3- Q1) 1,412.0…
    - Average spending in "Lisbon" is lesser compared to Oporto or Other, and average spend is highest in "Oporto" region, also spread (std) is highest in "Oporto" region.
    - Spends varies most in Hotels, but average spending is high in Retails.

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?

We are identifying inconsistent behaviour of variables (varieties) by Coefficient of Variation (CV)

Fresh	1.053918
Milk	1.273299
Grocery	1.195174
Frozen	1.580332
Detergents_Paper	1.654647
Delicatessen	1.849407

Figure F - CV of varieties

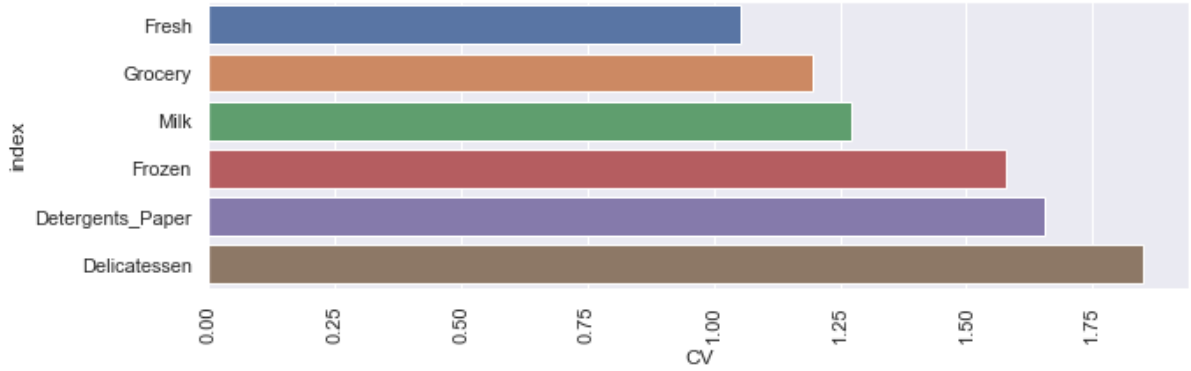

Figure G - Visual representation of CV

As we can see by above figures, we can observe that **Delicatessen** *(CV:1.849)* shows most inconsistent behaviour whereas **Fresh** *(CV: 1.053)* has least inconsistency.

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

Statistically, outlier are values that are not between Q1 – 1.5 IQR and Q3 + 1.5 IQR (*IQR – Inter-quartile range)

For our varieties, the IQR and respective thresholds are –

	IQR	Q1-1.5IQR	Q3+1.5IQR
Fresh	€ 13,806.00	-€ 17,581.25	€ 37,642.75
Milk	€ 5,657.25	-€ 6,952.88	€ 15,676.13
Grocery	€ 8,502.75	-€ 10,601.13	€ 23,409.88
Frozen	€ 2,812.00	-€ 3,475.75	€ 7,772.25
Detergents Paper	€ 3,665.25	-€ 5,241.13	€ 9,419.88
Delicatessen	€ 1,412.00	-€ 1,709.75	€ 3,938.25

Figure H - Tabular representation of Threshold deciding outliers

And the number of outliers we have in each variety is –

Fresh (20), Milk (28), Grocery (24), Frozen (43), Detergents Paper (30) and Delicatessen (has 27)

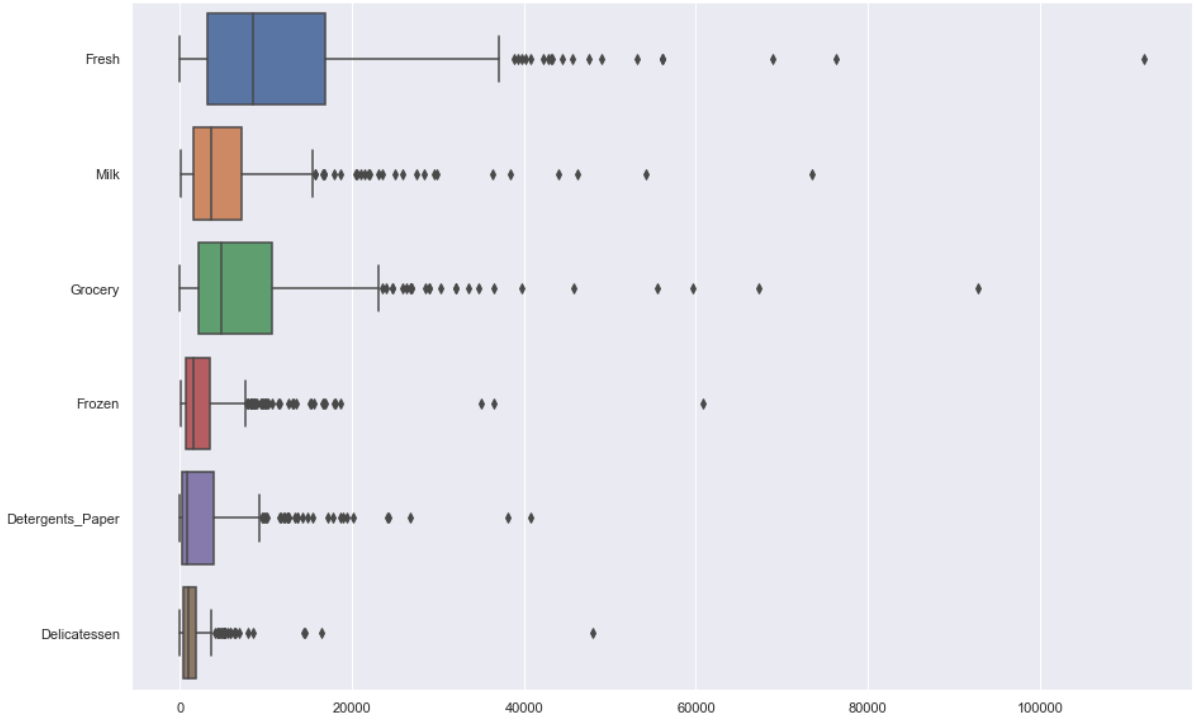

Figure I - Visual Representation of Outlier

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.

We have some observations on which business can take up some action to increase sales (spending).

- Oporto has less spending on Fresh, Milk & delicatessen, so business can focus and try to increase sales.
- Hotels have less spending on Milk, Groceries and detergent paper, business can place some offers to increase sales there.
- Delicatessen has more inconsistency in spending compared to others, which can be reduced by increasing the spend in this variety.
- There are extreme spends per verity, business can have more stores opened on places where people are spending high amount, to gain trust and attract more customers by reducing cost by some amount.

Problem – 2 (Survey)

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.

Basic Data Definition:

ID: Serial number of the records

Gender: Gender of student (Male/Female)

Age: Age of student

Class: Current class of student (Junior, Senior or Sophomore)

Major: Preferred major of student (Accounting, CIS, Economics/Finance, International Business, Management, Retailing/Marketing, Other, Undecided)

Grad Intention: Student's intention with respect to graduation (Yes, No or Undecided)

GPA: Secured grade points

Employment: Current employment status (part-time, full-time, or unemployed)

Salary: Salary of student

Social Networking: Number of ways student being connected with social networking (0-4)

Satisfaction: Satisfaction level (1-6)

Spending: Amount student is spending

Computer: Type of machine student owns (Laptop, Desktop or Tablet)

Text Messages: Number of text message student sends

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

Contingency (cross-tabulation) table would help us exploring two or more categorical variables.

2.1.1 Gender and Major

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

Figure J - Contingency table between Gender and Major

By above figure, we can infer that Retails/Marketing & Economics/Finance are common major people are preferring, and some males were not decided about their major.

2.1.2 Gender and Grad Intention

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

Figure K - Contingency table between Gender and Grad Intention

9 females don't have intention to pursue for graduation and 17 out of 29 males want to graduate.

2.1.3 Gender and Employment

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

Figure L - Contingency table between Gender and Employment

43 out of 62 students are part-time employed, 10 full-time and 9 are not employed.

2.1.4 Gender and Computer

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

Figure M - Contingency table between Gender and Computer

Most students own Laptop, and 2 female students uses Tablet.

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

By any contingency table in above question (Figure J), it is evident that we have 33 female students and 29 male students.

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

Probability of a random selected student would be male is:

Number of Males / Total Number of Students

# of Males = 29

# of Females = 33

Total Students = 62

P(Male) = 29/62 ➔ 46.77 %

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

P(Female) = 33/62 ➔ 53.22 %

Probability that a randomly selected student will be male is 46.77 % and probability that it will be female is 53.22 %

2.3 Assume that the sample is a 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.

By contingency table between Gender and Major (Figure J), we can calculate conditional probabilities of majors among the male students –

P (Accounting|Male)	4/29
P (CIS|Male)	1/29
P (Economics/Finance|Male)	4/29
P (International Business|Male)	2/29
P (Management|Male)	6/29
P (Other|Male)	4/29
P (Retailing/Marketing|Male)	5/29
P (Undecided|Male)	3/29

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

We can observe conditional probabilities among female students, using Figure J

P (Accounting|Female)	3/33
P (CIS|Female)	3/33
P (Economics/Finance|Female)	7/33
P (International Business|Female)	4/33
P (Management|Female)	4/33
P (Other|Female)	3/33
P (Retailing/Marketing|Female)	9/33
P (Undecided|Female)	0/33

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.

Contingency/Frequency table between Gender and Grad intention (Figure K) would tell us the probability –

$$P(A \cap B) = P(B)*P(A)$$

$P(A \cap B)$ – Randomly chosen student is male that intends to graduate

P (B) – Randomly chosen student to be a Male student ➔ 29/62

P (A) – Male student that is intends to graduate ➔ 17/29

P(A∩B) = (29/62 * 17/29) = 17/62 = **27.42 %** possibility that a randomly chosen student is male and intends to graduate.

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

We would refer to frequency table Gender and Computer (Figure M) –

P (B) – Randomly chosen student to be a female student ➔ 33/62

P (A) – Female student does NOT have a laptop ➔ 1 – (29/33) = 4/33

P(A∩B) = (33/62 * 4/33) = 4/62 = **6.45 %** possible that a randomly chosen female student does not have a laptop.

2.5. Assume that the sample is a 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 a full-time employment

Referring to contingency table between Gender and Employment, we can identify probability by –

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

P(A∪B) -- Probability of randomly chosen student to be a Male or has a full-time employment

P(A) -- Probability of randomly chosen student is a Male ➔ 29/62

P(B) -- Probability of someone having full-time employment ➔ 10/62

P(A∩B) -- Probability of Male student having full-time employment ➔ 7/62

P(A∪B) = (29/62 + 10/62 – 7/62)

= 32/62

= **51.61 %**

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

$$P(A \mid B) = P(A \cap B) / P(B)$$

P(A|B) – Majoring in international business or management, given that she has been chosen randomly

P(B) – Randomly chosen student is a female ➔ 33/62

P(A∩B) – Randomly chosen student is a female AND majoring in international business or management ➔ 4/62 + 4/62

P(A|B) = (8/62) / (33/62)

= 8/33

= **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 graduate intention and being female are independent events?

Pulled subset of data (removing graduation intend "Undecided"), we are left with 40 students as 22 of 62 were not decided.

Out of 40, 11 female and 17 male students intends to pursue graduation.

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Figure N - Contingency table (2x2) Gender and Grad Intention

For independent events, we should have equation P(A∩B) = P(A) · P(B) holds true…

P(A∩B) – Probability of female students having graduation intend

P(A) -- Probability of students having graduation intend

P(B) – Probability of randomly chosen student is female

From above contingency table, we can infer –

P(A∩B) = 11/20

P(A) = 28/40

P(B) = 20/40

It is evident that P(A∩B) ≠ P(A)*P(B), so we can say both events are **NOT** independent.

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?

Probability will be the ratio of as – Students having less than 3 GPA / Total Students

We have 17 students will less than 3 GPA, and have 62 total students…

So,

Probability of student having less than 3 GPA would be = 17/62 ➔ **27.41 %**

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

Flagged the students based on their salary,

EarnsMoreThan50:

When 50.0 or more than **TRUE** else **FALSE**

And built a frequency table for Gender and EarnsMoreThan50 -

EarnsMoreThan50	False	True	All
Gender			
Female	15	18	33
Male	15	14	29
All	30	32	62

Figure O - Contingency table between Gender and Salary>=50.0

P (Earn more than 50 | Male) = 14/62 = **22.58 %**

P (Earn more than 50 | Female) = 18/62 = **29.03 %**

2.8.1 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.

1. **GPA:**

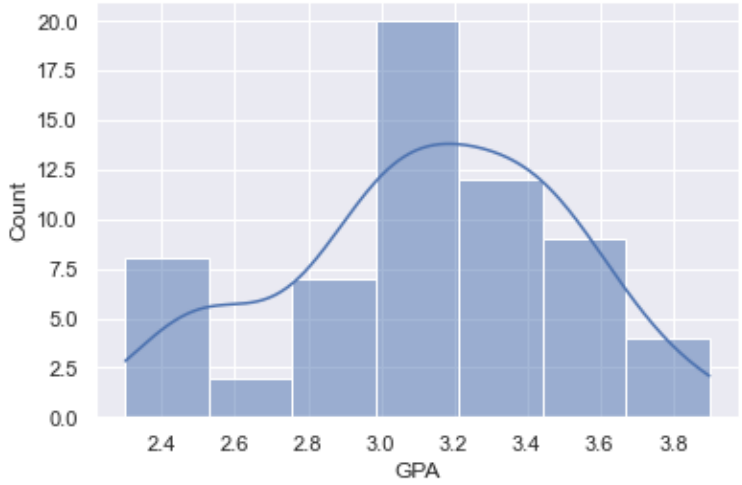

Figure P - Histogram of GPA

It has mean of 3.129, median 3.15, Mode 3.1 (nearby mean) and standard deviation of 0.377

- Mean, median and mode are not at one point, so we can say the probability curve is not bell-shaped curve but looks approximately normal distributed
- Empirical rule says 1 standard deviations (on each sides) from the mean would have 68% probability if data is normal distributed.

  $\mu = 3.129$
  $\sigma = 0.377$

  So, probability of GPA between $\mu+\sigma$ (3.506) and $\mu-\sigma$ (2.752) would be –

  # of students have GPA between 2.752 and 3.506 / total # of students

  = 45/62 = 72.58 % (Close to 68%)
- Probability of GPA at 1 std from mean is 72.58 %, at 2 std 96.77% and at 3 std probability will be 100%… It is evident that GPA **follows** normal distribution approximately.

2. **Salary:**

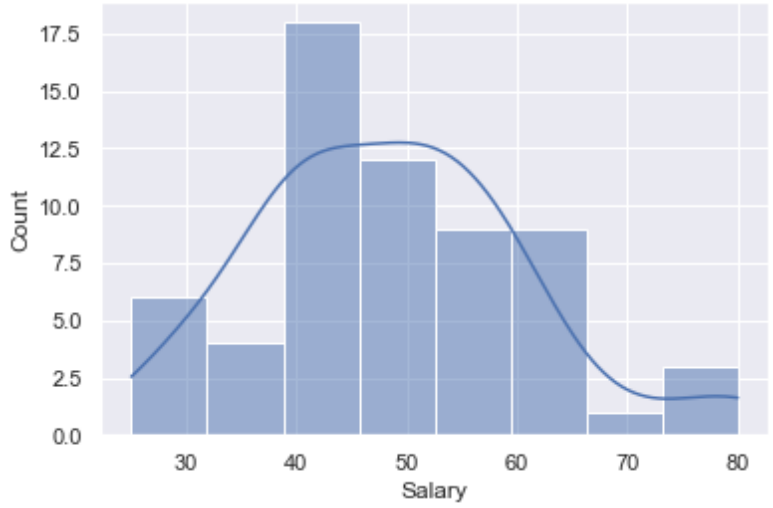

Figure Q - Histogram of Salary

It has mean of 48.548, median 50.0, Mode 40.0 (nearby mean) and standard deviation of 12.08

- Mean, median and mode are not at one point, so we can say the probability curve is not bell-shaped curve but looks approximately normal distributed
- Empirical rule says 1 standard deviations (on each sides) from the mean would have 68% probability if data is normal distributed.

  μ = 48.548
  σ = 12.08

- Probability of Salary at 1 std from mean is 79.03% (more than 68%), at 2 std 95.16% and at 3 std probability will be 100%… It is evident that salary **<u>does not follow</u>** normal distribution.

3. Spending:

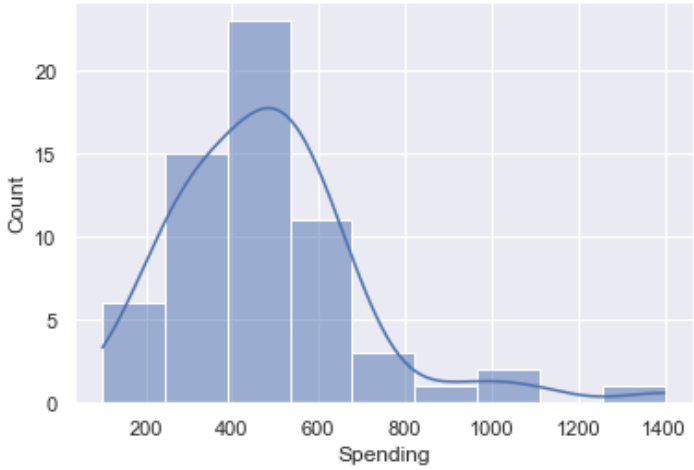

Figure R - Histogram of Spending

It has mean of 482.016, median 500.0, Mode 500.0 (nearby mean) and standard deviation of 221.95

- Mean, median and mode are not at one point, so we can say the probability curve is not bell-shaped curve but looks approximately normal distributed

- Empirical rule says 1 standard deviations (on each sides) from the mean would have 68% probability if data is normal distributed.

  μ = 482.016
  σ = 221.95

- Probability of Spending at 1 std from mean is 80.64% (more than 68%), at 2 std 95.16% and at 3 std probability will be 98.38%… It is evident that salary **does not follow** normal distribution.

4. Text Messages:

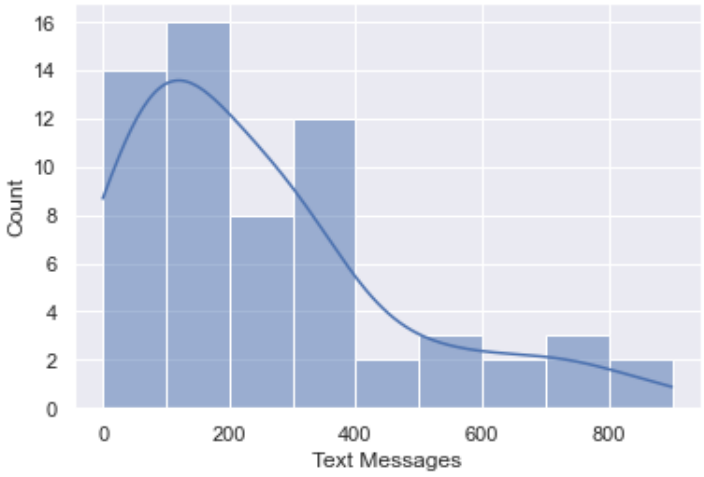

Figure S - Histogram of Text Messages

It has mean of 246.209, median 200.0, Mode 300.0 (nearby mean) and standard deviation of 214.46

- Mean, median and mode are not at one point, so we can say the probability curve is not bell-shaped curve but looks approximately normal distributed
- Empirical rule says 1 standard deviations (on each sides) from the mean would have 68% probability if data is normal distributed.

  μ = 246.209
  σ = 216.46

- Probability of Text messages at 1 std from mean is 79.03% (more than 68%), at 2 std 91.93% and at 3 std probability will be 98.38%… It is evident that salary **does not follow** normal distribution.

2.8.2 Write a note summarizing your conclusions

- **GPA** looks approximate normally distributed, but **Salary**, **Spending** and **Text Messages** have more data available in 1 standard deviation from the mean (much more than 68%), so we can infer that they don't follow normal distribution.
- Skewness Calculated by Fisher-Pearson correlation –

    - GPA has a value of -0.306 (Left skewed)
    - Salary has 0.521
    - Spending has 1.5472
    - Text Messages has 1.26
- We can conclude on skewness based on the histogram plots and above data –
    - **GPA** has median > mean, it means the data is left skewed and same is evident by histogram and Fisher-Pearson correlation.
    - **Salary** has median > mean, it means the data is left skewed, but data says otherwise, that means Salary data in not truly continuous, and outliers are present in it.
    - **Spending** has median > mean, it means the data is left skewed, but data says otherwise, so we can assume data in not continuous.
    - **Text Messages** has median < mean, it means the data is right skewed and same is evident by histogram and calculated values.

Problem 3 (A & B Shingles)

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.

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.

We will perform hypothesis testing on both types of Shingles to state if they are within the permissible limits…

Hypothesis testing on Shingle A

Stating NULL and Alternate hypothesis -

HA (Alternate hypothesis) ➔ Mean moisture content (μ) < 0.35

H0 (NULL hypothesis) ➔ Mean moisture content (μ) >= 0.35

Calculated statistical values from sample –

Mean of Sample ($\bar{x}$): 0.317

Standard Deviation of Sample (s): 0.136

# of items in Sample (n): 36

As we don't know the population standard deviation and only one sample (Shingle A) in question, we can perform 1-sample T-test to take decisions on our stated hypothesis.

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

Calculated T-statistics and P-value –

T-statistics = -1.473

P-value = 0.0747 (One – tailed test)

If we assume level of significance (α) as 0.05, it is evident that **P-value > α** and we are failed to reject NULL hypothesis…

***** We can conclude that mean moisture content of Shingle A is not under permissible limit (0.35)**

Hypothesis testing on Shingle B

Stating NULL and Alternate hypothesis -

HA (Alternate hypothesis) ➔ Mean moisture content (μ) < 0.35

H0 (NULL hypothesis) ➔ Mean moisture content (μ) >= 0.35

Calculated statistical values from sample –

Mean of Sample ($\bar{x}$): 0.274

Standard Deviation of Sample (s): 0.137

# of items in Sample (n): 31 (Including Blanks)

As we don't know the population standard deviation and only one sample (Shingle A) in question, we can perform 1-sample T-test to take decisions on our stated hypothesis.

Calculated T-statistics and P-value –

T-statistics = -3.10

P-value = 0.002 (One – tailed test)

If we assume level of significance (α) as 0.05, it is evident that **P-value < α** and we can reject NULL hypothesis with 95% confidence…

***** We can conclude that mean moisture content of Shingle B is under permissible limit (0.35)**

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?

Defining NULL and Alternate hypothesis –

µA – Population mean of shingle A

µB – Population mean of shingle B

HA (Alternate hypothesis) ➔ µA = µB

H0 (NULL hypothesis) ➔ µA ≠ µB

***This would a two-sampled T-test for the null hypothesis that 2 independent samples having equal population mean, the test assumes that both populations have **identical variance**…

- 2 – Tailed T-test of independent samples
- Level of significance (α) taken as 0.05

Calculated T-statistics and P-value:

T-statistics = 1.289

P-value = 0.201

It is evident that **P-value > α** , hence we are failed to reject NULL hypothesis… The population means of Shingle A and Shingle B are **not equal.**