Statistical Methods for Decision Making - Report

Contents

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

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: **- Page 7**

2.4.1. Find the probability That a randomly chosen student is a 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.

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: **- Page 8**

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

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

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? **- Page 8**

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. **- Page 9**

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

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.

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 for this whole Problem 2. **- Page 10**

Problem 3:

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.**- Page 12**

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? - **Page 13**

Problem 1: Wholesale Customers Analysis

Problem Statement:

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 channels (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?

1.1 Answer:

- The "Other" Region has spent the most.
- The "Hotel" Channel has spent the most.

Code Used:

```
df_1['Total'] = df_1['Fresh']+df_1['Milk']+df_1['Grocery']+df_1['Frozen']+df_1['Detergents_Paper']+df_1['Delicatessen']

df_1.groupby("Region").sum()

df_1.groupby("Channel").sum()
```

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.

1.2 Answer:

- Across all regions the "Other" Region has the most sales, followed by Lisbon.
- The Fresh Variety is most popular, followed by Grocery.
- In the Hotel channel - Fresh is the most popular variety, whereas in Retail - Delicatessen is the most popular category.

Code Used:

df_1.groupby("Region").sum().iloc[:,1:]

df_1.groupby("Channel").sum().iloc[:,1:]

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?

1.3 Answer:

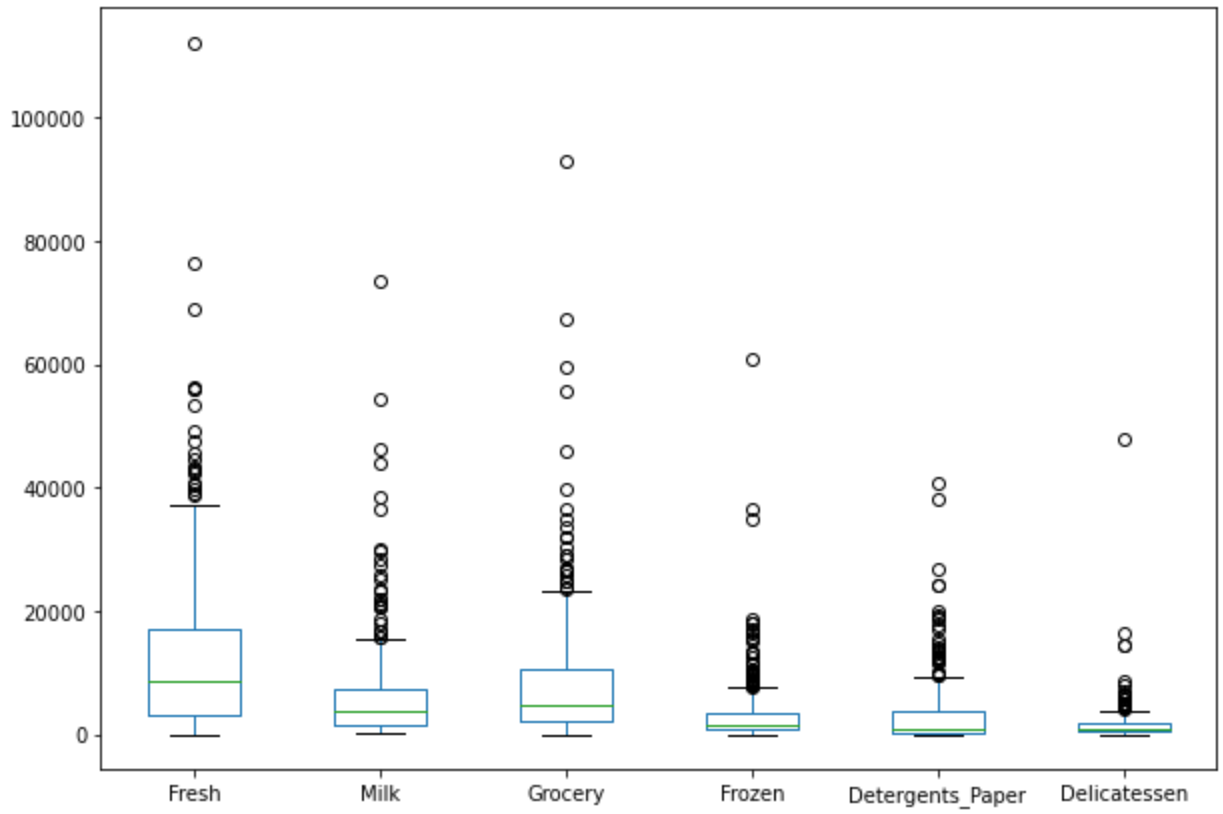
- Most inconsistent is - Delicatessen
- Least inconsistent is - Fresh

Code Used:

```
print("Coefficient of Variation of Fresh is "+str(stats.variation(df_1.iloc[:,1:]["Fresh"])))

print("Coefficient of Variation of Milk is "+str(stats.variation(df_1.iloc[:,1:]["Milk"])))

print("Coefficient of Variation of Grocery is "+str(stats.variation(df_1.iloc[:,1:]["Grocery"])))

print("Coefficient of Variation of Frozen is "+str(stats.variation(df_1.iloc[:,1:]["Frozen"])))

print("Coefficient of Variation of Detergents_Paper is "+str(stats.variation(df_1.iloc[:,1:]["Detergents_Paper"])))

print("Coefficient of Variation of Delicatessen is "+str(stats.variation(df_1.iloc[:,1:]["Delicatessen"])))
```

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

1.4 **Answer:**



It is very clearly visible from the boxplot that all the numerical columns have outliers.

Code Used:

df_1.iloc[:,1:9].plot(kind='box', figsize=(10,7))

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.

1.5 Answer:

- The Other Region and Hotel Channel are the maximum spenders. More supply centers should be built close to them.
- In the Hotel channel - Fresh is the most popular variety, whereas in Retail - Delicatessen is the most popular cateogry.In the respective channels we must focus on strengthening the supply even more.
- There is a lot of inconsistency in the behaviour, which is a negative sign. This should be corrected.

Problem 2: Clear Mountain State University (CMSU) Survey Analysis

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

2.1.1 Answer: 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

Code Used: gender_major = pd.crosstab(df_2['Gender'], df_2['Major'], margins = False)

gender_major = pd.crosstab(df_2['Gender'], df_2['Major'], margins = False)

2.1.2 Answer: Gender and Grad Intention

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

Code Used: gender_gradintent = pd.crosstab(df_2['Gender'], df_2['Grad Intention'], margins = False)

2.1.3 Answer: Gender and Employment

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

Code Used: gender_employment = pd.crosstab(df_2['Gender'], df_2['Employment'], margins = False)

2.1.4 Answer: Gender and Computer

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

Code Used: gender_comp = pd.crosstab(df_2['Gender'], df_2['Computer'], margins = False)

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?

2.2.1 Answer: 0.4677

Code Used: df_2[df_2.Gender=="Male"]["ID"].count()/df_2["ID"].count()

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

2.2.2 Answer: 0.532

Code Used: df_2[df_2.Gender=="Female"]["ID"].count()/df_2["ID"].count()

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.

2.3.1 Answer:

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Male	0.137931	0.034483	0.137931	0.068966	0.206897	0.137931	0.172414	0.103448

Code Used: gender_major.iloc[1:,:]/gender_major.iloc[1:,:].sum(axis=1)[0]

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

2.3.2 Answer:

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	0.090909	0.090909	0.212121	0.121212	0.121212	0.090909	0.272727	0.0

Code Used:

gender_major.iloc[0:1,:]/gender_major.iloc[0:1,:].sum(axis=1)[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.

2.4.1 Answer: 0.586

Code Used: gender_gradintent.iloc[1:,2:]/gender_gradintent.iloc[1:].sum(axis=1)[0]

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

2.4.2 Answer: 0.1212

Code Used: 1-gender_comp.iloc[0:1,1:2]/gender_comp.iloc[0:1].sum(axis=1)[0]

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?

2.5.1 Answer: 0.5161

Code Used: ((7+19+3)+(3+7)-(7))/(3+7+24+19+6+3)

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

2.5.2 Answer: 0.2424

Code Used: (4+4)/(3+3+7+4+4+3+9+0)

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?

2.6 Answer:

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Code Used:

gender_gradintent.drop(["Undecided"], axis =1)

Here let us consider A = female, B = Graduation Intention

P(A∩B)=11/40 P(A)=20/40 P(B)=28/40

P(A∩B)=11/40=0.275 P(A)·P(B) = 0.35

As they are not equal, it can be concluded that these events are NOT independent.

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?

2.7.1 Answer: 0.274

Code Used:

df_2[df_2["GPA"]<3].count()[0]/df_2.count()[0]

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.

2.7.2 Answer: Male: 0.3448 and Female: 0.3939 respectively.

Code Used:

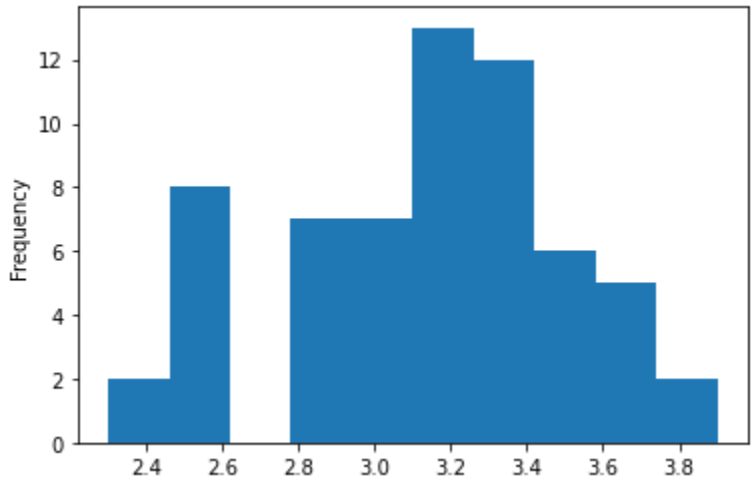
```
df_2[df_2["Gender"]=="Male"][df_2["Salary"]>50].count()[0]/df_2[df_2["Gender"]=="Male"].count()[0]
```

```
df_2[df_2["Gender"]=="Female"][df_2["Salary"]>50].count()[0]/df_2[df_2["Gender"]=="Female"].count()[0]
```

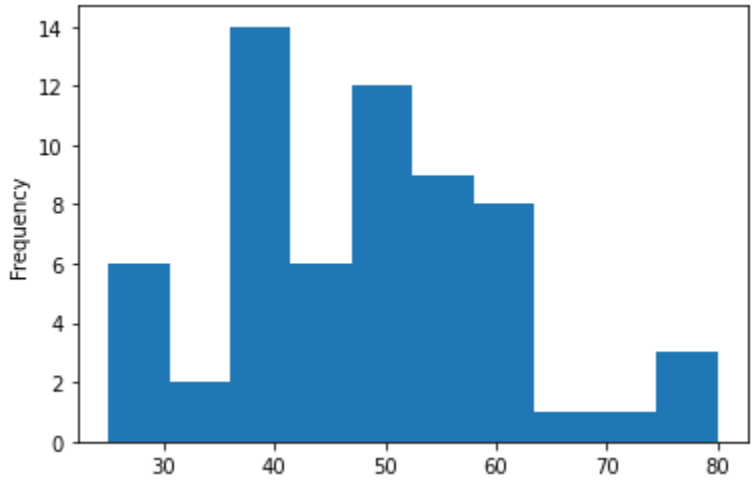
2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions for this whole Problem 2.

2.8 Answer:

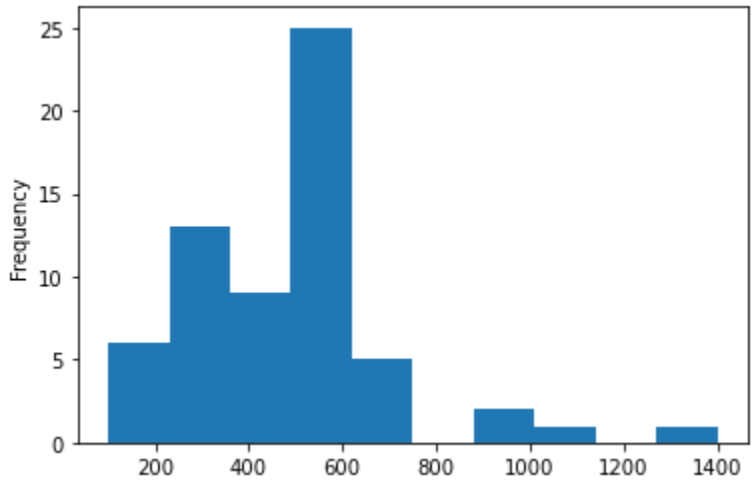
GPA:



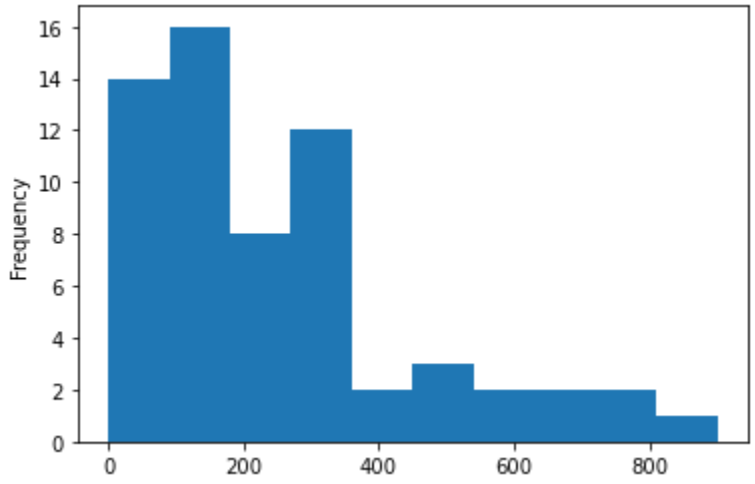
Salary:



Spending:



Text Messages:



It can be concluded that GPA and Salary are Normal Distributions, whereas Spending and Text Messages are not.

Overall the following conclusions can be made about the data:

- Female students constitute approximately 53.2% of the college.
- Most popular Majors for Male students are: Management and Retailing/Marketing.
- Most popular Majors for Male students are: Retailing/Marketing and Economics/Finance.
- CIS and Accounting are the least popular among Male and Female.
- Gender and Intent to Graduate are not independent events.

Code Used:

```
df_2["GPA"].plot.hist()

df_2["Salary"].plot.hist()

df_2["Spending"].plot.hist()

df_2["Text Messages"].plot.hist()
```

Problem 3: ABC Asphalt 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.

3.1 Answer:

Shingle A:

t statistic: -1.4735046253382782 p value: 0.07477633144907513

Inference: As p values > 0.05, we can't reject H0, there is NOT ENOUGH evidence to conclude: moisture contents in Sample A shingles are within the permissible limits

Shingle B:

t statistic: -3.1003313069986995 p value: 0.0020904774003191826

Inference: As p values < 0.05, we can reject H0, there is ENOUGH evidence to conclude: moisture contents in Sample B shingles are within the permissible limits

Code Used:

```
t_statistic, p_value = stats.ttest_1samp(df_3.A, 0.35)

print('t statistic: {0} p value: {1} '.format(t_statistic, p_value/2))

t_statistic, p_value = stats.ttest_1samp(df_3.B, 0.35,nan_policy='omit' )

print('t statistic: {0} p value: {1} '.format(t_statistic, p_value/2))
```

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?

3.2 Answer:

t_statistic=1.29 and pvalue=0.202

Inference: As the p value > α , we can't reject H0; and we can say that population Mean for Shingles A and B are equal.

Code Used:

```
t_statistic,p_value = stats.ttest_ind(df_3.A,df_3.B,equal_var=True ,nan_policy='omit')

print("t_statistic={} and pvalue={}".format(round(t_statistic,3),round(p_value,3)))
```