Problem 1

Given that the data given contains sales data across 6 different food items, creating a separate column to sum the total sales

1. **Descriptive Stats to Summarize data:**
    a. By adding a new Column which sums up the total sales cross all the 6 food items, gives a total sales volume number.
    b. The whole data set is grouped by Region and Channels separately to sort it further to Max and Min Sales across Regions/ Channels separately.
    c. The output is as below.

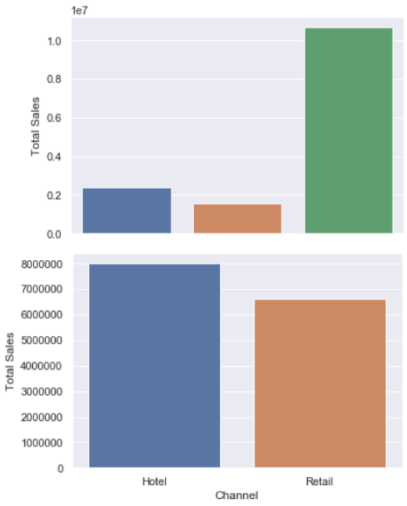
Sales Region-wise: **Maximum and Minimum**

Region	Total Sales
Other	**10677599**
Oporto	**1555088**

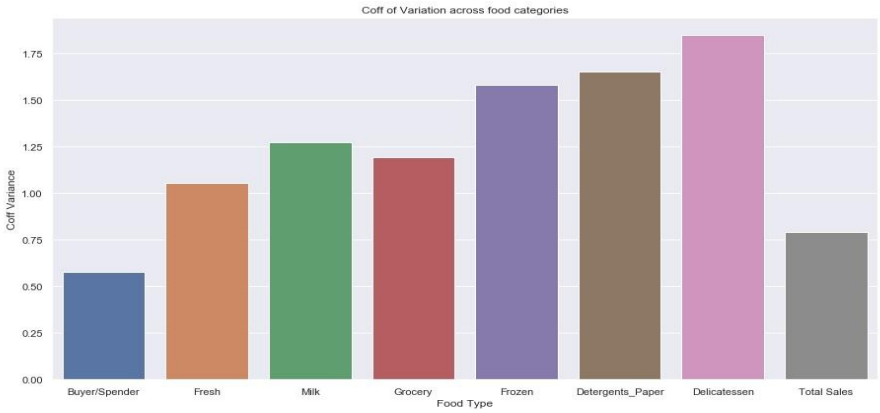
Sales Channel-wise: Maximum and Minimum

Channel	Total Sales
Hotel	**7999569**

Channel	Total Sales
Retail	**6619931**



2. **Do all the varieties show similar behaviour across Regions and Channels?**
    a. To assess and compare the measure of dispersion, we calculated Coffecient of Variance and added a new column.
    b. As per the bar plot of Coff of Var., there is no such similarity in the 6 different food items across the Regions and Channels.
    c. The output is as follows:



3. **Most Inconsistent v/s Lease Inconsistent behaviour in Food items.**
    a. One of the ways to measure the consistency is measured through comparing the Coff of Variance of various food types in the dataset.
    b. The one with the least Coff of Variance is the most Consistant one, and vice versa.
    c. The output is here under:

Inconsistent	Coff Variance
Delicatessen	1.84918
Consistent	**Coff Variance**
Fresh	**1.053917**

4. **Outliers in the data:**
    a. To determine any outliers in the data, Boxplots are best suited to find it out.
    b. The output shows that all the food categories have outliers. Please refer to the Jupyter file uploaded.

5. Recommendations: The company's strength is Hotel business. The number of Customers from Hotel channel outweighs the Retail by a significant margin. We grouped the data into Regions and then Channels to get a summation of various categories in different regions/ Channels.
    a. If the company is focussing to expand to newer cities/locations (major cities), Hotel channel must be it priority.
    b. The company should focus on aggressively growing its (Fresh, Milk and Grocery Segments) Retail Revenue, especially in the cities where it is already present.

    c. Retail in Oorto: Although Retail is soft in other regions, Opport Region supports Retail business despite having lesser footfall. The management must focus growing aggressively in Retail segment

Region	Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_ Paper	Delicatessen	Total Sales
Lisbon	Hotel	14026	761233	228342	237542	184512	56081	70632	1538342
	Retail	4069	93600	194112	332495	46514	148055	33695	848471
Oporto	Hotel	8988	326215	64519	123074	160861	13516	30965	719150
	Retail	5911	138506	174625	310200	29271	159795	23541	835938
Other	Hotel	48020	2928269	735753	820101	771606	165990	320358	5742077
	Retail	16006	1032308	1153006	1675150	158886	724420	191752	4935522

Problem No 2

1. Please refer to the Jupiter file submitted for Contingency Tables
2. Conditional Probability Calculations
    a. The probability that a randomly selected student is a MALE is: 0.47
    The probability that a randomly selected student is a FEMALE is: 0.53

    b. The CONDITIONAL PROBABILITY of different Majors for MALE students is:

Major	Conditional P Male
Accounting	0.14
CIS	0.03
Economics/Finance	0.14
International Business	0.07
Management	0.21
Other	0.14
Retailing/Marketing	0.17
Undecided	0.10

The CONDITIONAL PROBABILITY of different Majors for FEMALE students is

Major	Female
Accounting	0.09
CIS	0.09
Economics/Finance	0.21
International Business	0.12
Management	0.12
Other	0.09
Retailing/Marketing	0.27
Undecided	-

    c. The conditional probability of intent to graduate, given that the student is a male is: 0.58
    The conditional probability of intent to graduate, given that the student is a male is: 0.33

    d. The CONDITIONAL PROBABILITY of different EmploymentStatus for FEMALE students is as follows:

Employment	Female
Full-Time	0.09
Part-Time	0.73
Unemployed	0.18

The CONDITIONAL PROBABILITY of different EmploymentStatus for MALE students is as follows:

Employment	Male
Full-Time	0.24
Part-Time	0.66
Unemployed	0.10

e. The CONDITIONAL PROBABILITY of Laptop Preference for MALE students is: 0.9
The CONDITIONAL PROBABILITY of Laptop Preference for FEMALE students is: 0.88

3. The variables in the Columns are Dependent on the Gender.
    a. Reason no 1.: The probabilities is a function of number of Male/Females in the observation.
    b. Reason no2: The probabilities and conditional probabilities shall change as any change in number of Males or females in the observation.

4. To test the Normalcy of the data, for columns -> 'Salary', 'Spending' and 'Text Messages'. We tried to verify its normalcy through
    a. Shapiro test data.
        i. Salary-> The Shapiro test p value comes out to be 0.028

        The value is below the significance value of 0.05. Hence the observations could not be considered as normal

        ii. Spending-> The Shapiro Test p-value is = 1.68*10^(-5)
        The value is still far below the significance level of 0.05. Hence , the distribution could not be considered as normal.
        iii. Text Message -> The Shapiro Test p-value is = 4.32*10^(-6)
        The value is still far below the significance level of 0.05. Hence , the distribution could not be considered as normal.
    b. Boxplot Analysis: As per Boxplot of the 3 columns,
        i. Salary displays some normalcy
        ii. Spending and Text Messages does not.
    c. There are other ways also which can be used to check for normalcy, but that doesn't give a quantified answer.
        i. For example, check for Mean= Median in describe function. The closer the mean to median, the probability of the distribution to be normal is higher.
        ii. Also, plotting histogram also gives a visual of normalcy, but cannot be relied with certainity.

Problem 3:

3.1 What assumption do you need to check before the test for equality of means is performed?

Answer: The Assupmtion of Independent Samples are to be checked.

3.2 What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

Answer: Following assupmtions:

1. The assumption which is needed in order to carry out the t-test, is that the sample collected follows normal distribution.
2. The measurement applied to the data is of continuous scale.

This is the end of the report.