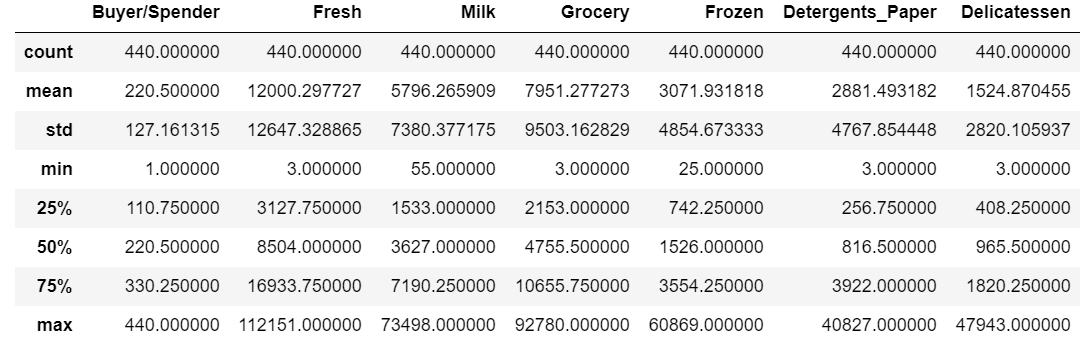
**Wholesale Customers Analysis**

**Problem Statement:**

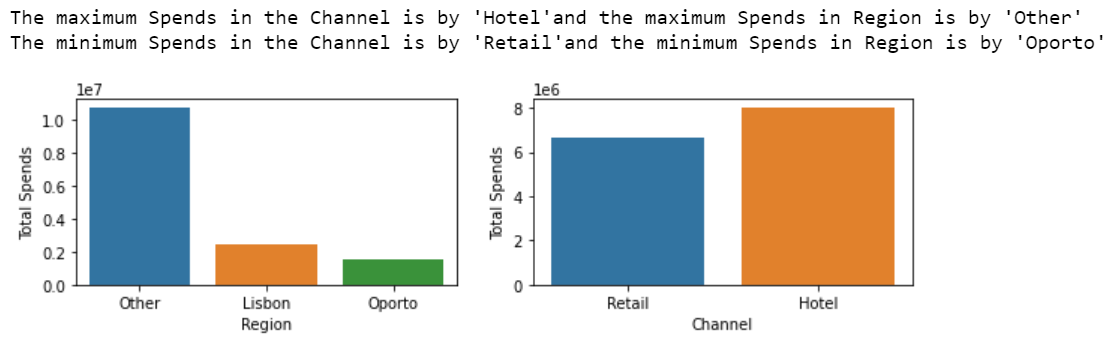
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers’ annual spending on 6 different varieties of products in 3 different regions (Lisbon,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

1. **Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?**

* In this case we will focusing on the descriptive **statistics as in the case for summarizing the data we will be using the describe () in-built function and the data which we get from python when we use describe function is something like:**



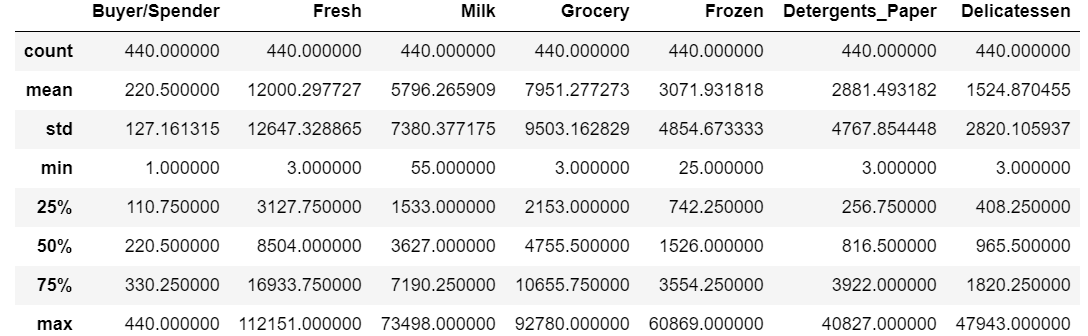
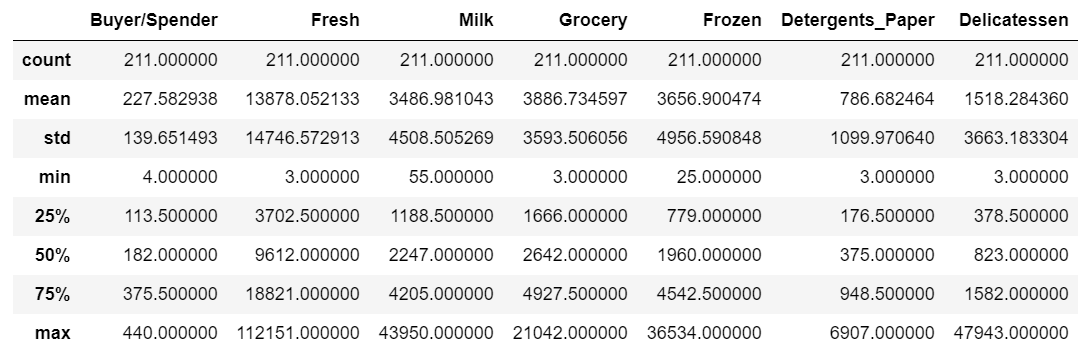
* Now talking about the second part of the question that is which Region and Channel spends more and less. Now in this case we will using the for loops which will calculate the total spends for all varieties according to the Region and Channel. As we can see from the data Region and Channel are categorical values so we have to calculate for each category and then visualize which category spends more and spends less. Once we are done with all the calculation, we plot it with the help of bar graph as we can visually get a idea which category present in Region and Channel Spends more or less. The graphical representation is as follows:



* As we can see in above representation, we can get an idea which category in region and channel spends more or less.

1. **There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?  Provide justification for your answer.**

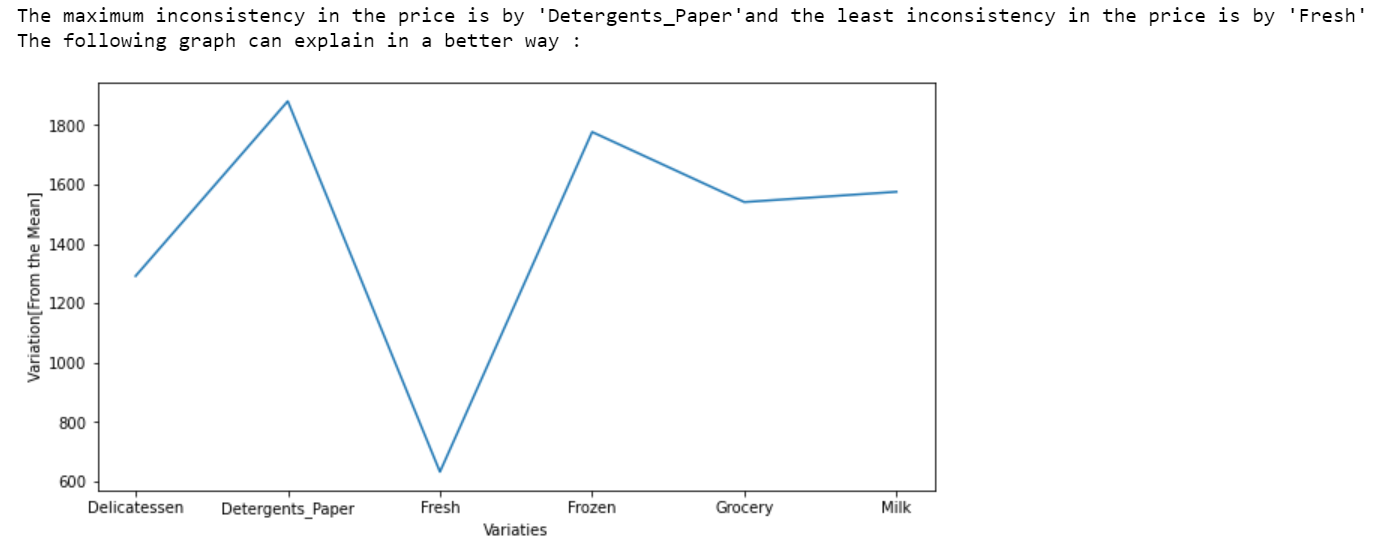
* Yes, the varieties show the same behaviour across Region and Channel. As we get insights from the data if we are taking any conditions regarding the category for example if Retail category is Other and Channel category is Hotel and when this is taken into consideration these categories when combined with each other can give us a great impact on the varieties. But if in general we talk regarding that if these varieties have impact across Retail and Channel as a whole then we might not get any impact as such.



* In the above table the data on which describe function is applied is by taking in consideration the Region category Other and Channel category Hotel. If we now compare it from the above table, we will be looking at a lot of difference in the data as categories are taken into consideration.

1. **On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**

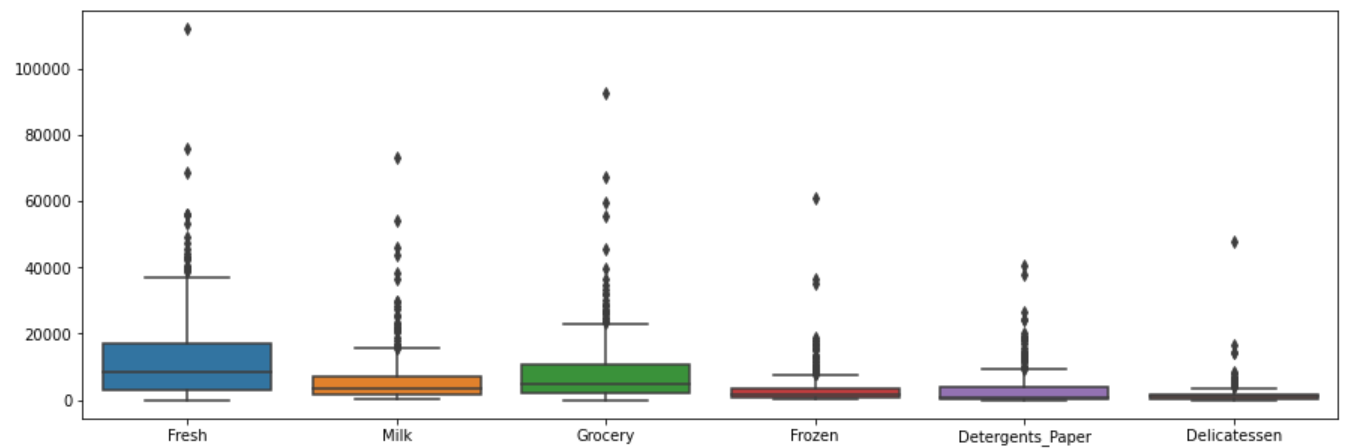
* Now as we know from the above answers, we can get an idea that **descriptive measure is always taken with the help of describe function.**
* Without spending much a time on it we will jump to the second part which implies that which varieties show the higher variance or inconsistency.
* For accomplishing this task, we have to look at some of the parameters which can help us identify the variance now that can be done with the help of IQR [Inter Quartile Range]. IQR formula is Q3[75% Quartile] – Q1[25% Quartile]. These quartile values are easily available from the descriptive table and can be calculated quite with ease.
* The second option, and this is the option which according to me makes more sense is the Standard Deviation. First, we calculate the standard Deviation once we do that, we get a set of 6 S.D values for each variety.
* These S.D values can be subtracted from the mean values of each variety so that we can understand how much variation is present in the data from the mean values.
* To understand this properly, a line plot will be a better option so that visualization becomes easy task. The line graph looks something like:



* In the above line graph, we can easily visualize that there is more inconsistency in the price of Detergents\_Paper and less inconsistency in the price of Fresh.

1. **Are there any outliers in the data?**

* To get the outliers in the data the best way is to generate a box plot as it is more informative to check the outliers visually rather than from the data as it becomes a task to analyse which data point is an outlier.
* Now with the help of box plot outliers which are present in the data set are:



* As we can analyse from above box plot each variety has an outlier value.

1. **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.**

* Now if according to business perspective we analyse the data we received; we can conclude from this data set that there are outliers present in each variety.
* The outliers are those values that can create a problem while analysis of the data is taken into consideration.
* There are two basic ways to deal with these types of data sets. First, either we have to remove the outliers present from the data set.
* But if we take the first method into consideration then we might end up in an improper or not properly distributed data.
* This can be a risky task as if the data is imbalance these values might be replaced with the NULL values and this will create a problem while visualizing the data.
* The Second method can be we can leave the outliers as it is but scale them according to the data set. For example, if we have data set like 1,2,3,4,5,1000. Now clearly in this case we can see that 1000 is an outlier now what can be done is that we can scale the value of 1000 in such a way that it does not create any major impacts on the model building supposing like 10. As we incur from other data points all the values are single digits so if we scale 1000 as 10 it will give us a proper mean and other parameters required and will not create major impact on the model building.

**Survey For CMSU**

**Problem Statement:**

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**

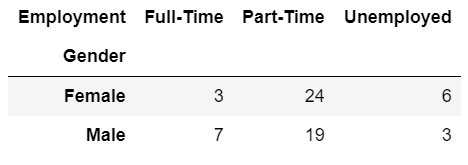
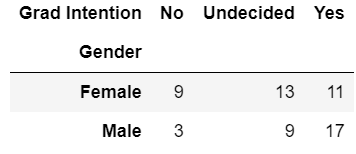
**2.1.2. Gender and Grad Intention**

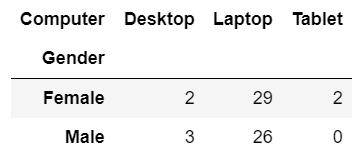
**2.1.3. Gender and Employment**

**2.1.4. Gender and Computer**

* For all the questions above same type of command will be used known as crosstab. This command returns us a matrix type data set which represents relation between the Gender and other Columns. The cross table for each data set looks something like this:







**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?**

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

* In the above tow questions, we are asked regarding the conditional probability of randomly selected student is a male and female.
* In this case we will be using the formula for **Conditional Probability** which is P (Students | Male) = P (Students which are male) / Total Students.
* The same formula applies for the randomly selected student is a female P (Students | Female) = P (Students which are female) / Total Students.
* This formula gives us the probability for male and female:

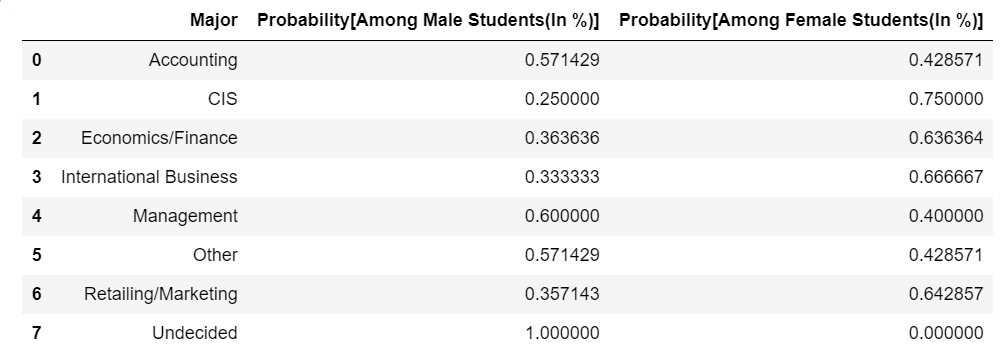


2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

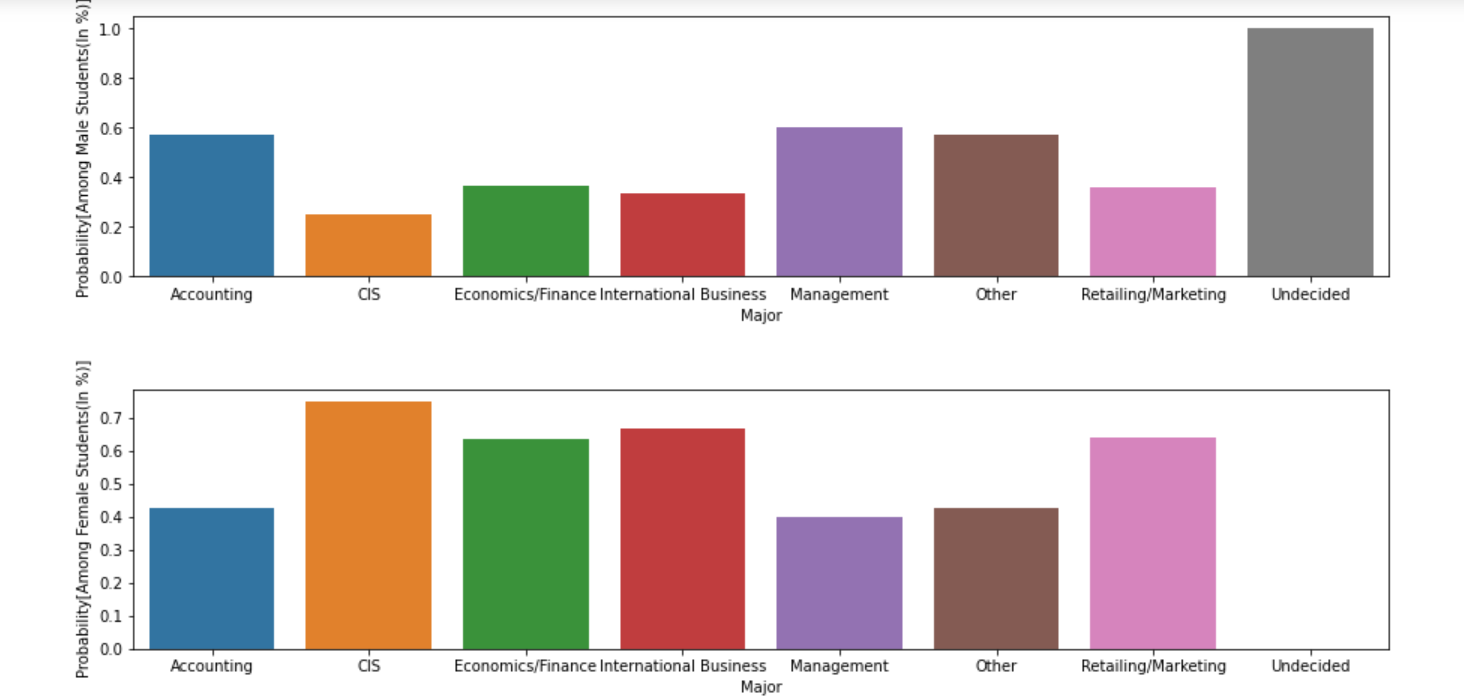
**2.3.1. Find the conditional probability of different majors among the male students in CMSU.**

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

* In the above question as well, we will be using conditional probability as we used in previous question but in this case we have to calculate the probability that what is the probability of male and female across different majors provided by CMSU.
* The calculated table with the probabilities for each major visually looks more appealing like this



* The graphical representation visually is represented as:

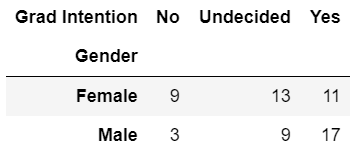


2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

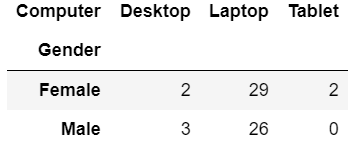
**2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.**

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

* In the first part we have check the graduation intention of male student which tends to graduate.
* In this scenario we have to calculate the conditional probability by using the cross table which we generated for gender and Grad Intent:



* In the table above as we can see there are 17 students who have an intention to graduate.
* The probability that the student intends to graduate is 17 / Total number of male students.
* In the second question we have to check for conditional probability for female student not having laptop.
* In this case as well we will be using the previously created cross table for gender and Computer.
* The same steps are applied for calculating conditional probability as well that is



* As we can observe or incur from the above table that there are 4 females not having laptops.
* Now for conditional probability we have to add females having desktop and Tablet as we have a condition that female does not have a laptop.
* The sum of female having desktop and tablet / Total number of females gives us the probability of Females not having Laptops.

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

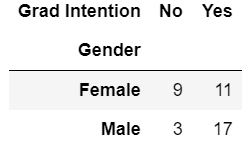
**2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment?**

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

* Now in the 2.5.1 we have to calculate the probability for separate conditions that is probability of student is male and probability that there is a full-time employment in this case we will be using a formula P (Male OR Full -Time Employment) = P (Male) + P (Full -Time Employment) – Intersection between Male and Full-Time Employment.
* In 2.5.2 we have to calculate the females which in majoring in international business and management.
* In this case the formula which will be applied will the probability of female with total students and the summation of females pursuing international business and 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?**

* The 2 X 2 table is represented by:



* Now when we calculate the graduation intention for female. Now in this case we will be calculating P (F ∩ Yes) = P(F)P(Yes)
* The Graduation Intention and Female are independent events.
* Why do we say so, we say this because we see for a fact that there is no relation between female and Graduation Intent.
* As we don’t have a previous data regarding if there has been any impact by any means which has changed the thought of a female regarding the Graduation. If for example we say that females which belong from lower class families not tend to graduate than the females who belong from the upper-class families. Then this might have an impact on the graduation as there are chances that lower class families might not be able to afford the education for their female child. If this would have been a situation then yes it would have been that female graduation depends on the intention of the graduation.
* This is the verdict which we all can relate to.

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

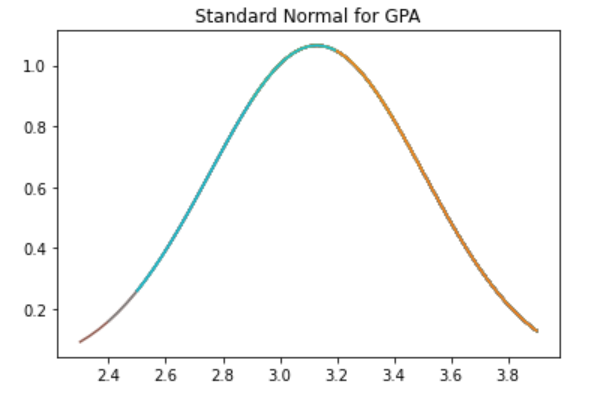
**2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?**

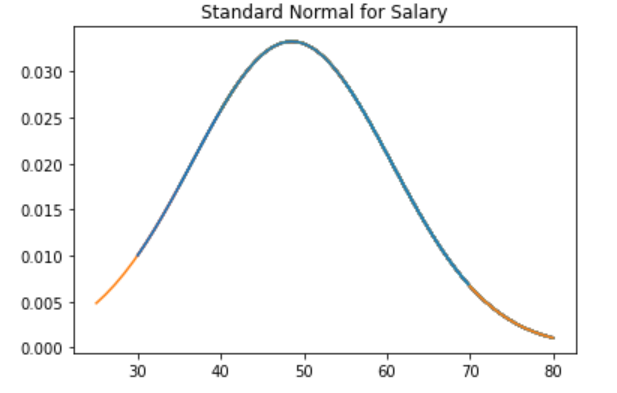
**2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.**

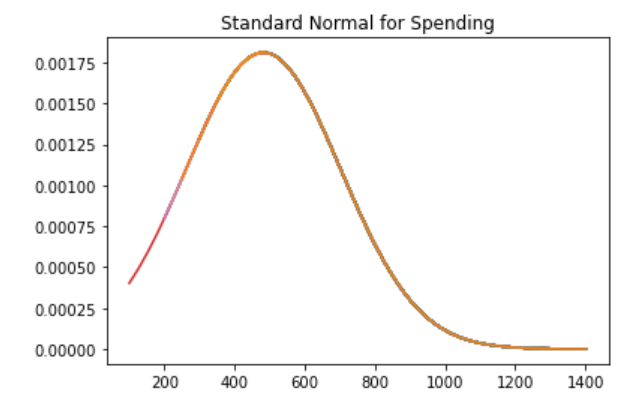
* Now to calculate the 2.7.1 we just have to use conditional probability the same way implemented in the previous questions.
* Just Divide the total students having GPA < 3 with the Total Students.
* In the question 2.7.2 we have to use the conditional probability for the males and females earning 50 or more.

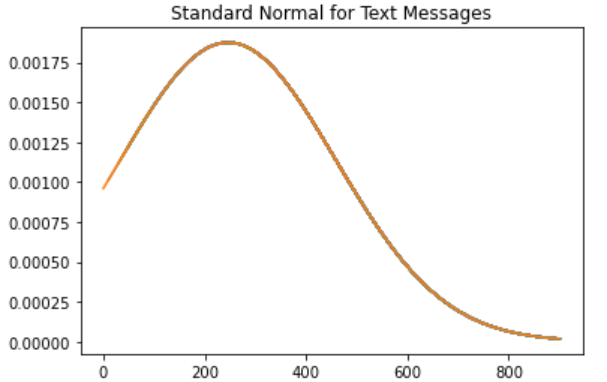
**2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions.**

* Now for checking for the normal distribution we have to visualize the data with the help of bell curve. If the data is normally distributed, we will have the mean, median and mode of the data almost the same which clearly indicates that the data is normally distributed.
* The bell curves for each continuous data is as follows:

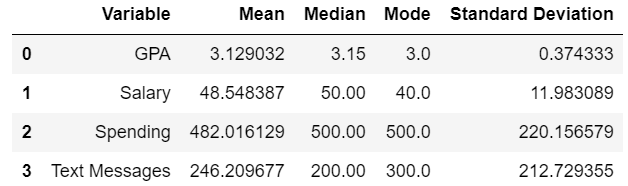








* This is the plots for each continuous data set we have.
* Now as we can incur from the data set, we observe that GPA and Salary have a normally distributed bell curve as compared to Spending and Text Messages which have a right tailed curve.
* Now, if we observe the table having mean, median and mode.



* We observe that the GPA and Salary has almost same values of mean, median and mode. This is a way to identify that if a continuous variable is normally distributed or not.
* Now coming to the steps followed. First, we take all the continuous variables calculate their mean, median, mode and standard deviation.
* Then, we have to calculate linspace which is a built-in command in the NumPy package.
* That linspace is applied to the plt plot which is also an inbuilt command in the matplotlib package.
* In this plt plot we have to provide the continuous variable and with that we have to use the stats.norm.pdf which is also an inbuilt function in the statistics package. In this we have to apply the continuous variable, mean and standard deviation.
* After all these calculations we can visualize a bell curve.

**Shingles**

**Problem Statement:**

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 pound per 100 square feet.

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

* In the above question we will be using the 1 sample ttest.
* We will be calculating the t statistics value and p-value for Shingles A and B separately to check whether the moisture content in both shingles are within the permissible limits.
* For checking this we will be focussing on the p-value and alpha value(**α).** By default, alpha value will be taken as 0.05 or 5%.
* If the p-value calculated in both A and B are greater than **α** then we can conclude that the moisture content is within the permissible limits or vice versa.
* As we incur from the data set, we conclude the fact that it is a one tailed test. So, for that we have to divide the p-value by 2.
* As by default, in python the p-value calculated from ttest\_1samp shows the result for 2 sided.
* Now when we calculate the p-value from A and B we observe that the p-value of A is 0.07477 > alpha (0.05) indicating that for A the moisture content in within the permissible limits.
* For B the p-value is 0.0020 < alpha (0.05) indicating that for B the moisture content is not within the permissible limits.

**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?**

* Step1: The Hypothesis formed will be

H0: **μ** < 0.35 [Null Hypothesis]

HA: **μ** > 0.35 [Alternate Hypothesis]

* Step2: Will be taking the alpha value by default as 5% or 0.05. As we are not been provided with the level of significance (**α).** So, if the value is not provided by default alpha is considered as 5%.
* Step3: Checking for which method can be best suitable for calculating the p-value. In this case we will be going for the 2sample T-Test(ttest\_ind).
* In python we have the command in the SciPy. Stats package for calculating 2sample T-Test. The command is scipy. stats. ttest\_ind.
* Which will return us a p-value. As mentioned earlier if the p-value is greater than alpha then we fail to reject Null hypotheses(H0).
* If the p-value is less than alpha we reject the Null Hypothesis meaning that Alternate Hypothesis is to be considered.
* Step4: Now as we have calculated the p-value we have come to the conclusion that p- value > alpha and hence we fail to reject null hypothesis.
* Which indicates that the mean moisture content which is claimed by the company to be less than 0.35 pounds per 100 square feet is true and the alternate hypothesis is false.