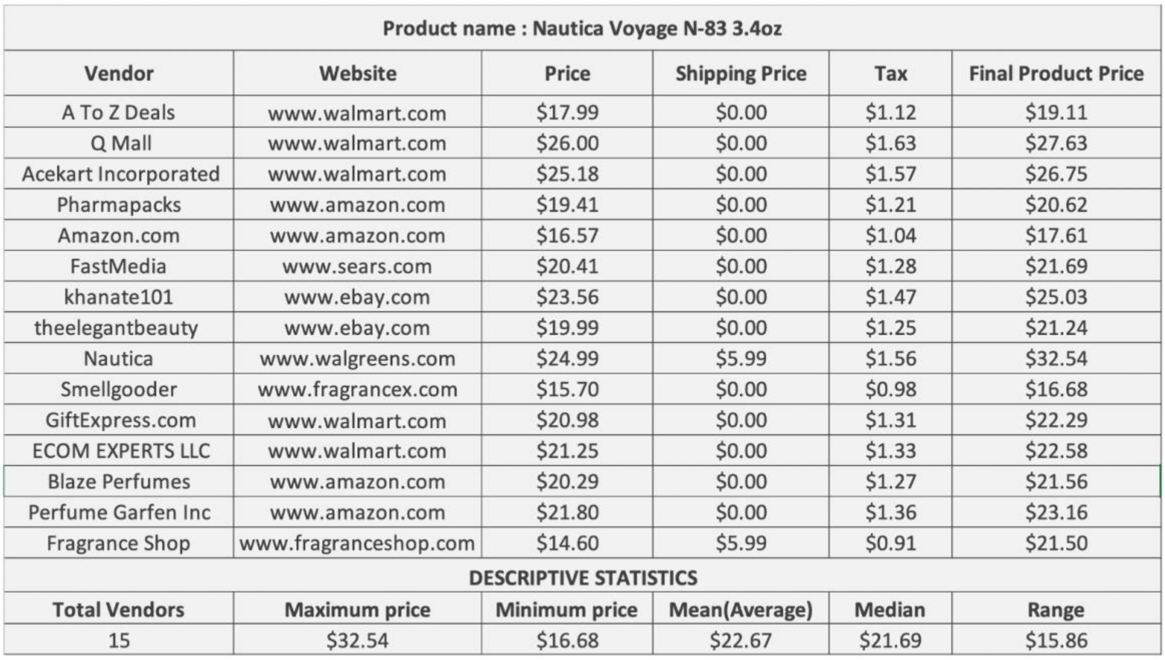


Product name : Nautica Voyage N-83 3.4oz					
Vendor	Website	Price	Shipping Price	Tax	Final Product Price
A To Z Deals	www.walmart.com	$17.99	$0.00	$1.12	$19.11
Q Mall	www.walmart.com	$26.00	$0.00	$1.63	$27.63
Acekart Incorporated	www.walmart.com	$25.18	$0.00	$1.57	$26.75
Pharmapacks	www.amazon.com	$19.41	$0.00	$1.21	$20.62
Amazon.com	www.amazon.com	$16.57	$0.00	$1.04	$17.61
FastMedia	www.sears.com	$20.41	$0.00	$1.28	$21.69
khanate101	www.ebay.com	$23.56	$0.00	$1.47	$25.03
theelegantbeauty	www.ebay.com	$19.99	$0.00	$1.25	$21.24
Nautica	www.walgreens.com	$24.99	$5.99	$1.56	$32.54
Smellgooder	www.fragrancex.com	$15.70	$0.00	$0.98	$16.68
GiftExpress.com	www.walmart.com	$20.98	$0.00	$1.31	$22.29
ECOM EXPERTS LLC	www.walmart.com	$21.25	$0.00	$1.33	$22.58
Blaze Perfumes	www.amazon.com	$20.29	$0.00	$1.27	$21.56
Perfume Garfen Inc	www.amazon.com	$21.80	$0.00	$1.36	$23.16
Fragrance Shop	www.fragranceshop.com	$14.60	$5.99	$0.91	$21.50
DESCRIPTIVE STATISTICS					
Total Vendors	Maximum price	Minimum price	Mean(Average)	Median	Range
15	$32.54	$16.68	$22.67	$21.69	$15.86

Did you have any difficulties when gathering your data?

Data collection was the fundamental part of the task. The very first step was shortlisting the product that needs to be explored for availability at various websites. Once I finalized the product, I felt there were a plethora of options available for the product on different websites. However, some vendors offered the product with a combo in turn lure users to buy products that may not be a necessity at that moment. Some of the vendors suggested products with a different variant, as the requested one was out of stock, again falls under the category of ‘not a necessity. This led me to explore more online sellers and different websites as well. Summing up the above points, I felt the challenges in the data collection procedure I experienced were the availability of the right product without unnecessary deals.

Which method of sampling did you use?

While sampling data for calculation I used Simple random sampling that falls under the Probability sampling technique. Each vendor that was used in the sample had equal chances of being selected. Each of the vendors was selected at random and had an equal number of chances of getting selected.

Explain how you choose the 15 vendors to survey and why you think they were chosen at random. Expound extensively, if possible.

Once the product was finalized the google search for the product returned around 100+ results for the availability of the product at different websites and different vendors. At first, when I selected a particular website for example Amazon, the price of the product offered by vendor1 was $X. I made a list of websites and vendors that were selling the required product. Next, I choose the vendors randomly from the list. There were around 5 other vendors on the same website that matched the requirement. At any point in time, there was no shortlisting or categorizing the vendors based on their location, reviews/ratings, or a number of years as a vendor on Amazon or any other parameter. Likewise, I followed the same procedure while looking into other websites as well. I also selected some websites from subsequent search pages apart from the initial results pages that google returned when the search for the product was made. This gave every vendor/website an equal probability of getting selected and hence the selection was random.

What inferences can you make about the structure of pricing ranges from the data and statistical analysis performed? Use your analytical skills.

Out of the 15 vendors where the data has been collected the price of the product ranges from $16.68 to $32.54 which has a range of $16. The average final price of the product available online was around $23. Around 67% of the vendors had prices less than the average price of the product. There would be more chances where customers buying the product would be paying around $23 or less.

```
dataset <- read.csv("/Users/lokaraju/ALY6000 - Course 1/Week 2/Discussion/D
iscussion2Data.csv", header = TRUE)

str(dataset)
```

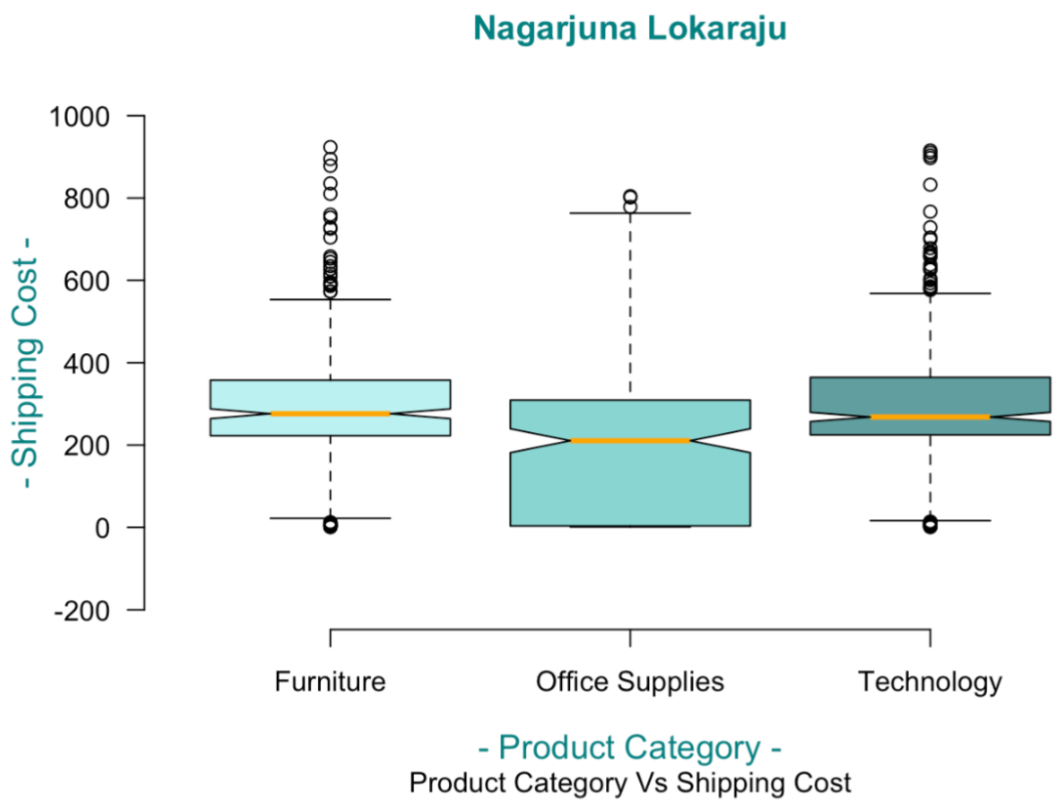
```
## 'data.frame':    1000 obs. of  10 variables:

##  $ Region              : chr  "Central US" "Oceania" "Oceania" "Western E
urope" ...

##  $ Market              : chr  "USCA" "Asia Pacific" "Asia Pacific" "Europ
e" ...

##  $ Company_Segment     : chr  "Consumer" "Corporate" "Consumer" "Home Off
ice" ...

##  $ Product_Category    : chr  "Technology" "Furniture" "Technology" "Tech
nology" ...

##  $ Product_SubCategory: chr  "Phones" "Chairs" "Phones" "Phones" ...

##  $ Sales               : chr  "221.98" "3,709.40" "5,175.17" "2,892.51" .
..

##  $ Quantity            : int  2 9 9 5 8 5 4 6 2 1 ...

##  $ Total_Sales         : chr  "443.96" "33,384.60" "46,576.53" "14,462.55
" ...

##  $ Profits             : chr  "62.15" "-288.77" "919.97" "-96.54" ...

##  $ ShippingCost        : num  40.8 923.6 915.5 910.2 903 ...
```

```
  Mean <- mean(dataset$ShippingCost)

Median <- median(dataset$ShippingCost)

Standard_Deviation <- sd(dataset$ShippingCost)

Minimum_Value <- min(dataset$ShippingCost)

Maximum_Value <- max(dataset$ShippingCost)

Range <- range(dataset$ShippingCost)

Quantile25 <- quantile(dataset$ShippingCost, 0.25)

Quantile50 <- quantile(dataset$ShippingCost, 0.50)

Quantile75 <- quantile(dataset$ShippingCost, 0.75)

library(kableExtra)


Particulars <- c('Mean', 'Median', 'Standard_Deviation', 'Minimum_Value', '
Maximum_Value', 'Range', 'Quantile25', 'Quantile50', 'Quantile75')

Values <- c("272.39", "258.90", "176.16", "1.07", "923.63", "922.56", "209.
82", "258.90", "351.07")

Table1 <- data.frame(Particulars, Values)


Table1 %>%

  knitr::kable(caption = "<br>**Nagarjuna Lokaraju**", color = "black") %>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "bordered", "hover", "condensed", "responsive", stripe_color = "blue"), full_width = FALSE, position = "left", html_font = "sans")%>%

  gsub("font-size: initial !important;", "font-size: 45pt !important;", .)%>%

  column_spec(2, bold = T, color = "teal", width = "1.5in")%>%

  column_spec(1, bold = T, color = "#515151", width = "1.5in")%>%

  row_spec(0, color = "white", background = "#008080")
```

Nagarjuna Lokaraju

Particulars	Values
Mean	272.39
Median	258.90
Standard_Deviation	176.16
Minimum_Value	1.07
Maximum_Value	923.63
Range	922.56
Quantile25	209.82
Quantile50	258.90
Quantile75	351.07

```
boxplot(dataset$ShippingCost ~ dataset$Product_Category, main="Nagarjuna Lokaraju", xlab = "- Product Category -", ylab = "- Shipping Cost -", col = c("#BBF1F1", "#89D5D2", "#5f9ea0"), ylim = range(-200:1000), col.lab = "#008080", notch = TRUE, frame.plot = FALSE, box.wex = 0.4, las = 1, medcol = "orange", cex.lab = 1.2, col.main = "#008080", cex.main = 1.2, sub = "Product Category Vs Shipping Cost")
```



```
dataset <- read.csv("/Users/lokaraju/ALY6000 - Course 1/Week 2/Discussion/D
iscussion2Data.csv", header = TRUE)

str(dataset)
```

```
## 'data.frame':    1000 obs. of  10 variables:

##  $ Region            : chr  "Central US" "Oceania" "Oceania" "Western E
urope" ...

##  $ Market            : chr  "USCA" "Asia Pacific" "Asia Pacific" "Europ
e" ...

##  $ Company_Segment   : chr  "Consumer" "Corporate" "Consumer" "Home Off
ice" ...

##  $ Product_Category  : chr  "Technology" "Furniture" "Technology" "Tech
nology" ...

##  $ Product_SubCategory: chr  "Phones" "Chairs" "Phones" "Phones" ...

##  $ Sales             : chr  "221.98" "3,709.40" "5,175.17" "2,892.51" .
..

##  $ Quantity          : int  2 9 9 5 8 5 4 6 2 1 ...

##  $ Total_Sales       : chr  "443.96" "33,384.60" "46,576.53" "14,462.55
" ...

##  $ Profits           : chr  "62.15" "-288.77" "919.97" "-96.54" ...

##  $ ShippingCost      : num  40.8 923.6 915.5 910.2 903 ...
```

```
  Mean <- mean(dataset$ShippingCost)
```

```r
Median <- median(dataset$ShippingCost)
Standard_Deviation <- sd(dataset$ShippingCost)
Minimum_Value <- min(dataset$ShippingCost)
Maximum_Value <- max(dataset$ShippingCost)
Range <- range(dataset$ShippingCost)
Quantile25 <- quantile(dataset$ShippingCost, 0.25)
Quantile50 <- quantile(dataset$ShippingCost, 0.50)
Quantile75 <- quantile(dataset$ShippingCost, 0.75)
library(kableExtra)

Particulars <- c('Mean', 'Median', 'Standard_Deviation', 'Minimum_Value', 'Maximum_Value', 'Range', 'Quantile25', 'Quantile50', 'Quantile75')
Values <- c("272.39", "258.90", "176.16", "1.07", "923.63", "922.56", "209.82", "258.90", "351.07")
Table1 <- data.frame(Particulars, Values)

Table1 %>%
  knitr::kable(caption = "<br>**Nagarjuna Lokaraju**", color = "black") %>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "bordered", "hover", "condensed", "responsive", stripe_color = "blue"), full_width = FALSE, position = "left", html_font = "sans")%>%
  gsub("font-size: initial !important;", "font-size: 45pt !important;", .)%>%
  column_spec(2, bold = T, color = "teal", width = "1.5in")%>%
  column_spec(1, bold = T, color = "#515151", width = "1.5in")%>%
  row_spec(0, color = "white", background = "#008080")
```

```r
boxplot(dataset$ShippingCost ~ dataset$Product_Category, main="Nagarjuna Lokaraju", xlab = "- Product Category -", ylab = "- Shipping Cost -", col = c("#BBF1F1", "#89D5D2", "#5f9ea0"), ylim = range(-200:1000), col.lab = "#008080", notch = TRUE, frame.plot = FALSE, box.wex = 0.4, las = 1, medcol = "orange", cex.lab = 1.2, col.main = "#008080", cex.main = 1.2, sub = "Product Category Vs Shipping Cost")
```

Probability

Give a succinct, understandable summary of your findings.

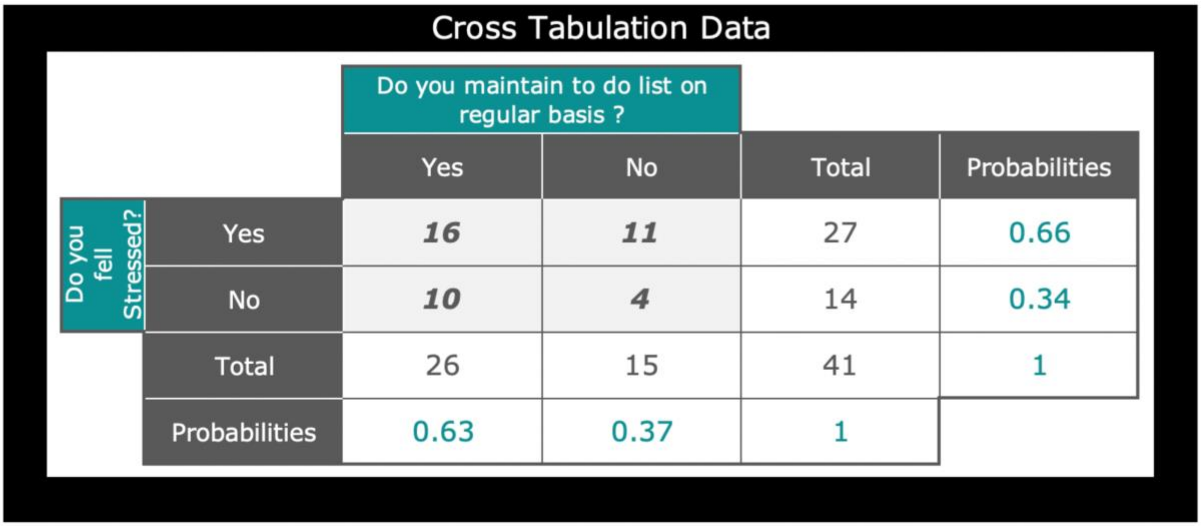
My intention of this research is to find what percent of people feel stressed when they maintain the to do list. I have collected the data from posting questions in the discussion panel and through google forms with single selection option and single response acceptance from one email id. Total of 41 responses are received. I found that 27 people are stressed and 26 people maintain to do list, which is the major observation.

Keep a record of your questions in the precise order that you ask them to your subjects.

1. Do you maintain to do list on a regular basis?
2. Do you feel stressed ?

Question :If a person maintains a to do list, what is the probability that he/she is stressed?

Answer: From the table we can see that 26 people who maintain the to do list on daily basis, out of which 16 are stressed. Hence the probability of people maintaining to do list and getting stressed is 16/20 = 0.80



Cross Tabulation Data

		Do you maintain to do list on regular basis ?			
		Yes	No	Total	Probabilities
Do you fell Stressed?	Yes	***16***	***11***	27	0.66
	No	***10***	***4***	14	0.34
	Total	26	15	41	1
	Probabilities	0.63	0.37	1	

Question: If a person is stressed, what is the probability that he/she is not maintaining to do list

Answer: Here, from the data, we can see that 27 members are feeling stressed, out of which 11 are not maintain the to do list. Hence the probability of stressed person not maintaining to do list is 11/27 = 0.40

Compile all of your findings into a comprehensive report.

As per the data received and analysed, we see that there is a direct proportional relationship between the maintenance of to do list and feeling stressed. This might be due to the to-do list is making them be overcommitted to the tasks and for which they are taking stress to complete. Reduce the list of tasks and being realistic as to what they can complete in each day can help them be out of stress.

The whole process is a good which starts from posting the questions and responding to all in the discussion. To achieve the required sample I had to reach out to friends and got to know the insights. Preparing the contingency table. Framing questions as to know what are the possible outcomes of the data.

2. Bayes' Theorem

Bayes’ theorem also knows as Bayes’ rule is a formula which can help in finding out the event probability based on the other related event. This theorem describes the possibility of the events happening, based on knowledge on the factors which are relevant to the event.

Bayes’ theorem – P(A|B) = P(B|A) x P(A) / P(B)

P(A|B) = Probability of event A occurring, considering event B has occurred.

P(B|A) = Probability of event B occurring, considering event A has occurred.

P(A). = Probability of event A

P(B) = Probability of event B

Application of Bayes theorem is into different industries such as Medicine, Pharma, Finance. Im interested in Finance industry where the application of Bayes theorem is all over, such as i) Probability of stock price changing due to change in interest rates, ii) Probability of change in selling price due to covid 19.

Here is the application in Finance industry – Using Bayes' theorem in Product Pricing.

30% of the companies that decreased their product selling price by more than 20% during the previous covid waves.

At the same time only 35% of the companies haven't decreased their product selling price by more than 20% during the covid waves.

Knowing that the probability that the product selling price decrease by more than 20% is 5%.

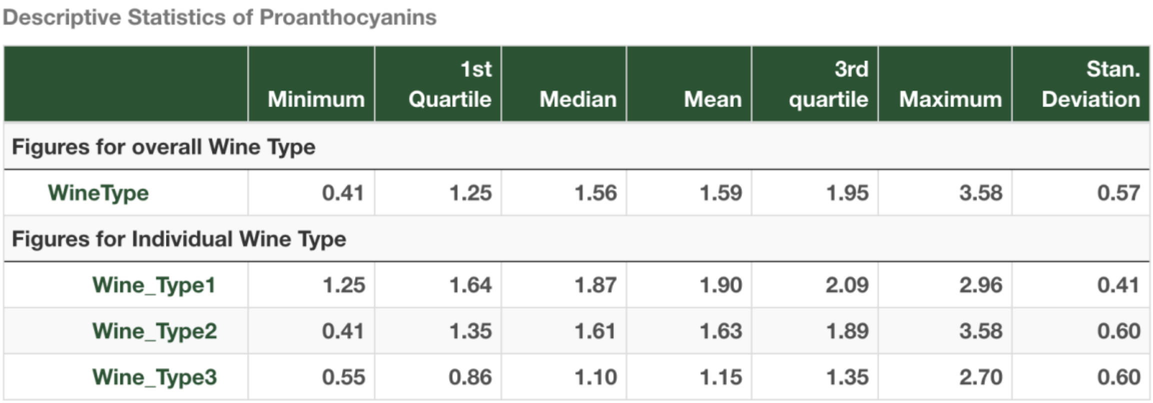
- P(A) – the probability that the product selling price decrease by 20%
- P(B) – the probability that there is a next covid wave
- P(A|B) – the probability of the product selling price decrease by 20% during the next covid wave has arrived
- P(B|A) – the probability of the Covid wave given the product selling price has decreased by 20%.

P(A|B) = 0.30 x 0.05 / (0.30 x 0.05) +0.35 x (1-0.05)

Using Bayes' theorem we probability that the product selling price of a company that will decrease by more than 20% during the next covid wave is 0.431

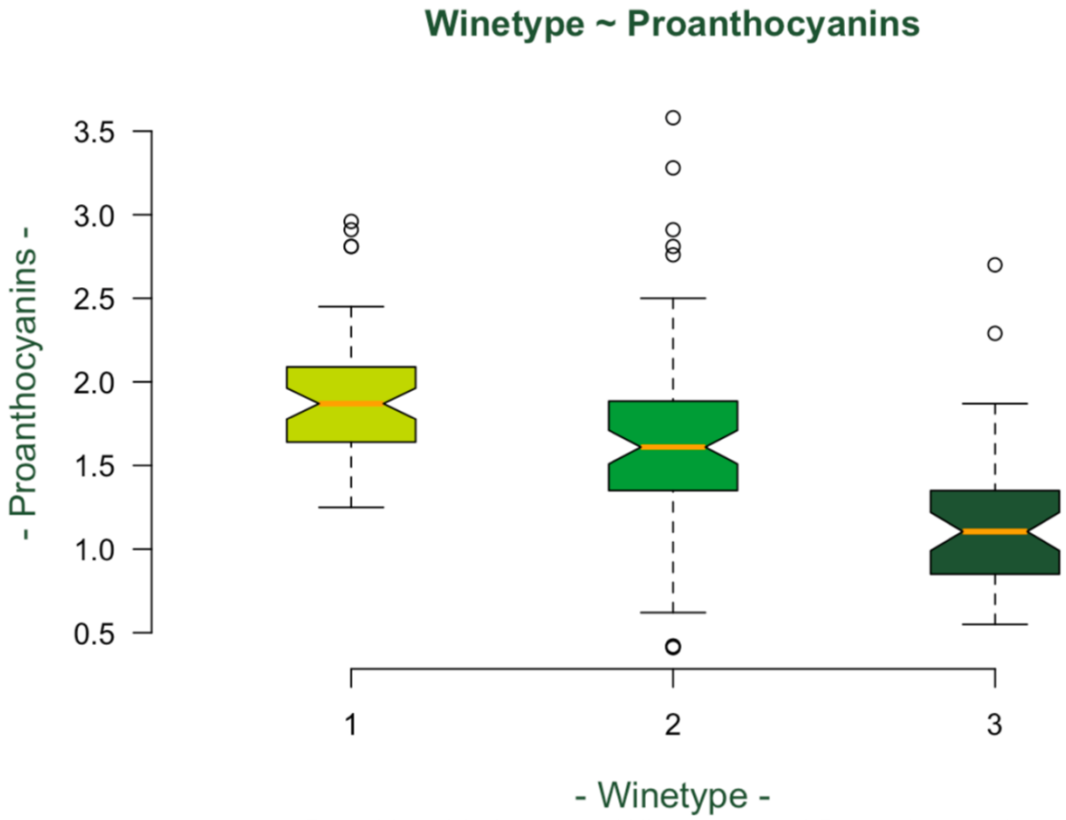
Part 1.1 Data Analysis

Numerical Value: ***Proanthocynins***

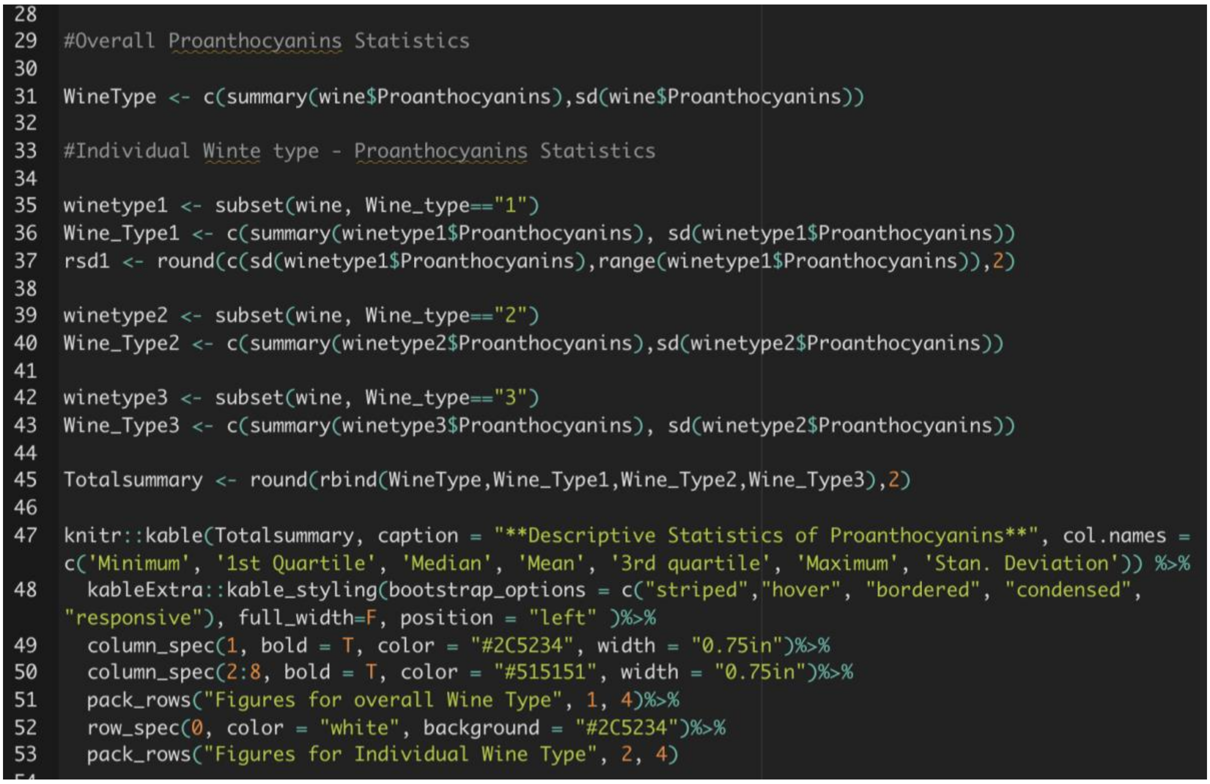


Descriptive Statistics of Proanthocyanins

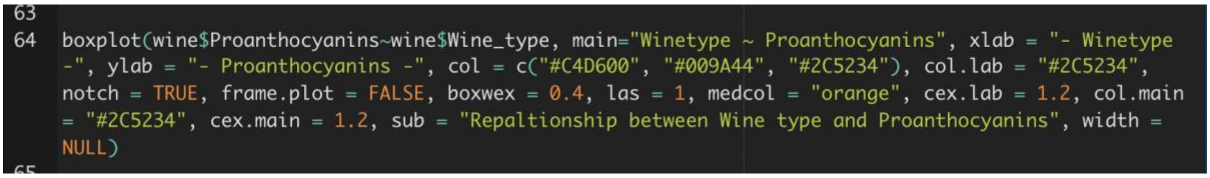
	Minimum	1st Quartile	Median	Mean	3rd quartile	Maximum	Stan. Deviation
Figures for overall Wine Type							
WineType	0.41	1.25	1.56	1.59	1.95	3.58	0.57
Figures for Individual Wine Type							
Wine_Type1	1.25	1.64	1.87	1.90	2.09	2.96	0.41
Wine_Type2	0.41	1.35	1.61	1.63	1.89	3.58	0.60
Wine_Type3	0.55	0.86	1.10	1.15	1.35	2.70	0.60



Repaltionship between Wine type and Proanthocyanins



```

#Overall Proanthocyanins Statistics

WineType <- c(summary(wine$Proanthocyanins),sd(wine$Proanthocyanins))

#Individual Winte type - Proanthocyanins Statistics

winetype1 <- subset(wine, Wine_type=="1")
Wine_Type1 <- c(summary(winetype1$Proanthocyanins), sd(winetype1$Proanthocyanins))
rsd1 <- round(c(sd(winetype1$Proanthocyanins),range(winetype1$Proanthocyanins)),2)

winetype2 <- subset(wine, Wine_type=="2")
Wine_Type2 <- c(summary(winetype2$Proanthocyanins),sd(winetype2$Proanthocyanins))

winetype3 <- subset(wine, Wine_type=="3")
Wine_Type3 <- c(summary(winetype3$Proanthocyanins), sd(winetype2$Proanthocyanins))

Totalsummary <- round(rbind(WineType,Wine_Type1,Wine_Type2,Wine_Type3),2)

knitr::kable(Totalsummary, caption = "**Descriptive Statistics of Proanthocyanins**", col.names = c('Minimum', '1st Quartile', 'Median', 'Mean', '3rd quartile', 'Maximum', 'Stan. Deviation')) %>%
  kableExtra::kable_styling(bootstrap_options = c("striped","hover", "bordered", "condensed", "responsive"), full_width=F, position = "left" )%>%
  column_spec(1, bold = T, color = "#2C5234", width = "0.75in")%>%
  column_spec(2:8, bold = T, color = "#515151", width = "0.75in")%>%
  pack_rows("Figures for overall Wine Type", 1, 4)%>%
  row_spec(0, color = "white", background = "#2C5234")%>%
  pack_rows("Figures for Individual Wine Type", 2, 4)
```



```
boxplot(wine$Proanthocyanins~wine$Wine_type, main="Winetype ~ Proanthocyanins", xlab = "- Winetype
-", ylab = "- Proanthocyanins -", col = c("#C4D600", "#009A44", "#2C5234"), col.lab = "#2C5234",
notch = TRUE, frame.plot = FALSE, boxwex = 0.4, las = 1, medcol = "orange", cex.lab = 1.2, col.main
= "#2C5234", cex.main = 1.2, sub = "Repaltionship between Wine type and Proanthocyanins", width =
NULL)
```

Observations

1. Proanthocyanin's quantity is highest in Wine type 1 and falling to the lowest quantity in Wine type 3
2. Concentration range of Proanthocyanin's is highest in Wine type 2, which is equal to the range of concentrations of all the wine types combined.
3. In wine type 1, the concentration range is lower, but the concentration is high

Conclusions

1. This graph can help in making decisions as to how the quality and concentration can make the difference of taste of wine
2. It can give insights into which type of grape is used for each wine type and what are its Proanthocyanins contents. Any modifications to be made in respect to the concentration and its taste aspects can be taken care of.

Part 1.2

1. Main characteristic of discrete probability distributions

The discrete probability distribution has two main characteristics

1. The probability of an event happening is always between 0 and 1 and inclusive of both (1)
2. The sum of all the probabilities of an event is 1

Example: Coin flip - two times

Outcomes = HH, HT, TH, TT

Considering the main objective as getting heads,

The probabilities are as follows

0 heads – 1 / 4 = 0.25

1 head – 2 / 4 = 0.50

2 heads – 1 / 4 = 0.25

Hence for the given example we can see that, Probability is always between 0 and 1 and the sum of (0.25+0.50+0.25) probabilities is 1.

2. Difference between discrete probability distribution and normal distributions

Discrete Probability Distribution	Normal Distributions
Finite outcomes	Any possible outcomes
Discrete	Continuous
Expressed in Tabular form	Expressed with Equation or Formula
Can obtain a particular value	Cannot obtain a particular value

3. Practical example of discrete probability distributions.

Under the Binomial Distribution concept which is one of the discrete probability distributions, outcomes are classified as either successful or a failure. (1)

In the game of cricket, the probability of a batsmen hitting six runs for every ball for continuous six deliveries, considering all the other things remain constant.

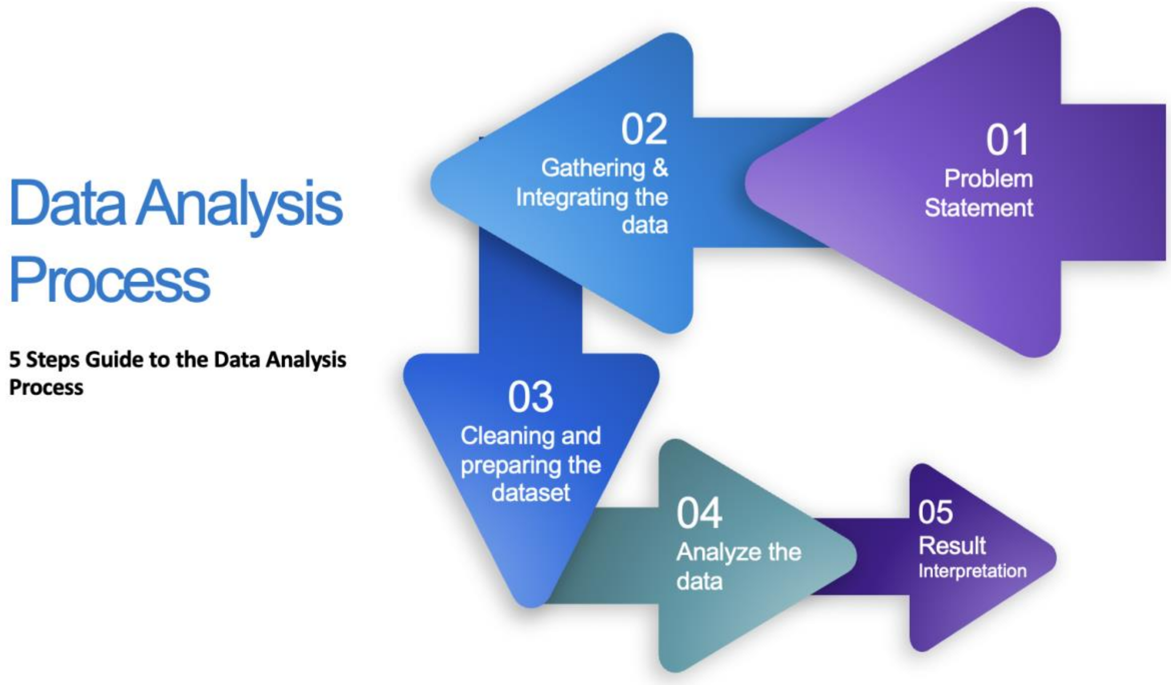
All possible events

1. Hitting a 6 = True
2. Not hitting a six = False

Probability = event occurrence/ all possible events

Every ball has a possibility of True or False = 1/ 2

Therefore, the probability of batsmen hitting six for all six balls is (1/ 2) ^ 6 = 1/ 64 => 0.0156.



Data Analysis Process

5 Steps Guide to the Data Analysis Process

01
Problem Statement
02
Gathering & Integrating the data
03
Cleaning and preparing the dataset
04
Analyze the data
05
Result Interpretation

Problem Statement

Many analysts are unsure as to what is the actual business issue and what are the steps to solve it. Brief and comprehensive description of an issue is very important, as it provides a clear aim to find the solutions (3). I believe outlining the challenges faced, will help the organization in finding the critical issues which impact the business and help them to work towards it. Therefore, defining the problem forms part of the basic step for data analysis.

Gather & Integrate Data

Not just defining the problem, but we also need to have data backing up for finding solutions. I feel that for the effective analysis and to deal with the issue, we need to have relevant data which should be collected. Hence the strategy for collecting the data with the right parameters such as timeframe from various sources be it direct / indirect source, structured/unstructured forms a crucial part.

Cleaning and preparing the Data Set

Data cleansing will improve overall productivity. Even though we have all the required data for analysis, from my perspective, there is a high possibility that the data might be affected with typos, automatically correcting words, Omission and commission errors, duplications, irrelevant, redundancy, outliers etc. Consequently, curating the data by removing all such possible errors (intentional or unintentional), noise will improve the quality, accuracy and make it as effective as possible.

Performing basic investigation on the data also will help you in improving the data. These can be done using R / Python / SQL and prepare the final output for analysis.

Analysing the data

The prepared dataset shall be applied with multiple statistical techniques (1) and alternate methods to extract meaningful insights to find solutions to the problem statement. I understand that all the above processes combined with analysis will help us to make data driven decisions which are far more effective than a guessing scenario (2). Analysis can be descriptive, predictive, perspective, diagnostic. Calculations/ processes followed should be right and accurate for the results to be correct, else the possibility of losing the authenticity of the analysis and the trustworthiness of the analysis will be lost.

Result interpretation

After all the steps with defining the issue, cleansing the collected right data for passing through different analysis and presenting the data using visualization tools, it's time for interpreting the results. If either of the steps above were not properly performed, interpretation might raise many questions. Analysis gives the summary of the data and its presentation factor; however, it is not the interpretation. Interpretation will give rise to new info, insights, key findings (1). All this will lead to a proper conclusion and solutions where actionable decisions will be made. It can also help in predicting the future outcomes and ideas to assess it.

B) Practical example of Data Analysis process

Losing high value SAP implementation deal (> $30M) to competitor, which is a must win deal of a strategic client.

Why have we lost?

Due to technical constraints and low operation capabilities – No,

Due to the high price quotation – yes (expected)

Why was it quoted high? because the partner went on gut feeling that the prices are good but not on the data backed decisions.

Problem statement: Loss of client due to wrong price quotation.

Required data

- Extract Sales pipeline and what are the upcoming engagements.
- Extract data pertaining to deals won and strategic clients from the past from the database
- Data pertaining to the lost deals of similar sort from the past.
- Competitor pricing information for lost deals. (third party sources), Knowledge exchange teams and research teams.

Data Cleansing

- Segregating all the deals which are of value range closer to $30M ($20M - $45M) to the opportunity lost, because lower value deals less or more than the range have different dynamics.
- Filtering the data based on time frames such as a year or two prior to this deal, as the prices and external business factors were not the same many years ago.
- Filtering data, which is similar to such an implementation, because other implementations would need different parameters to analyze.
- Removing duplicate sales entries and errors of such sort

Analysing the Data

- Using R we can process the data and find the descriptive analytics for such deals and other calculated fields such as weighted averages etc., as required.
- Make sure to check that all the calculations are right.
- Visualize the data with the timeline graphs to check the trends of pricing.
- Visualize the data as box plots to check where the actual numbers are falling
- Visualize the data with scatterplots to see similar kinds of deals clustered.

Interpretation

- After plotting our current deal metrics on the derived scatterplot, box plot, and timeline graph, we can see whether the price quotation made is matching with the metrics made as per analysis
- If the price quotation was not made in line with the analysis, then we can try implementing this for the upcoming deals and see if we are able to win the deal
- If the price quotation was made approximately matching with the analysis and still, we have lost the deal, we have to introspect into other questionable items where we can start our further analysis.