1. Pre-Project Plan

Goal Setting
I aim to complete my project by *14/1/2022* I shall take initiative to find out the information needed. I shall check the project rubric to ensure all items are done before submission.

Commented [LSL4]: Enter your target completion date. Set a week early will give you times to check incase certain task took longer than expected.

My data set is *e-Commercse_SetA*

My preliminary questions that I will answer from my data set (about 5):

Commented [LSL5]: Give at least 5 questions that you want to find out from the data.

1. *Which mode of flight has the highest demand?*

2. *Is there any correlation between the customer care call and customer rating?*

3. *Which variables have strong correlation with "reached on time"?*

4. *What is the customer rating by Gender?*

5. *What is the probability of delivery "reached on time"?*

6. Do customers with higher number of prior purchases had their delivery "reached on time" most of the time?

Project Monitoring

Commented [LSL6]: Jot down your actual completion date when you completed each task. This will help you to ensure your report can be submitted on time.

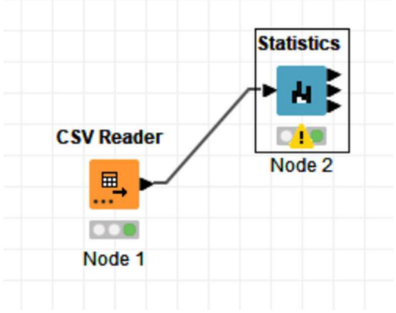
Task/Milestone	By When	Actual Completed Date	Comment (on-time/delay/early)
Download the data. Understand the rows and columns.	**21 Nov 2021**		
Background research of delivery mode, function of eCommerce.	**28 Nov 2021**		
Perform data cleaning.	**3 Dec 2021**		
Perform data transformation.	**8 Dec 2021**		
Exploratory Data Analysis	**17 Dec 2021**		
Submit Report 1	**24 Dec 2021**		
Answer my preliminary questions	**9 Jan 2022**		
Data modeling	**10 Jan 2022**		
Final report conclusion and reflection	**14 Jan 2022**		
Create Dashboard	**17 Jan 2022**		

3. Introduction

Many people are buying products online. There is a growing trend in eCommerce and the business is getting more and more competitive with more and more products selling on online platform.[1]

Being able to deliver product on time is an important service quality to stay competitive. This data contains information on shipment mode, customer care calls and customer rating. The company aims to gather some insights to check on the performance of the delivery and improve customer satisfaction.

Statistics node is used to learn more about the variables in the data.



File Edit Hilite Navigation View

Table "default" - Rows: 5501 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	S Warehouse_block	S Mode_of_Shipment	I Customer_care_call
61	D	Ship	3
62	F	Ship	3
63	A	Ship	5
64	B	Ship	3

There are 5501 observations.
11 Variables

Name of variable	*Data type*
Warehouse_block	*Categorical, A,B,C,D,E,F*
Mode_of_Shipment	*Categorical, Ship, Road, Flight*
Customer_care_call	*Numerical, discrete, between 2 to 7.* *This is obtained from the Statistics node*
Customer_rating	*Numerical, discrete between 1 to 5*
Cost_of-product	*Numerical, continuous, range from 93 to 310*
Prior_purchase	*Numerical, discrete range between 2 to 10*
:	
:	
:	

Continue …

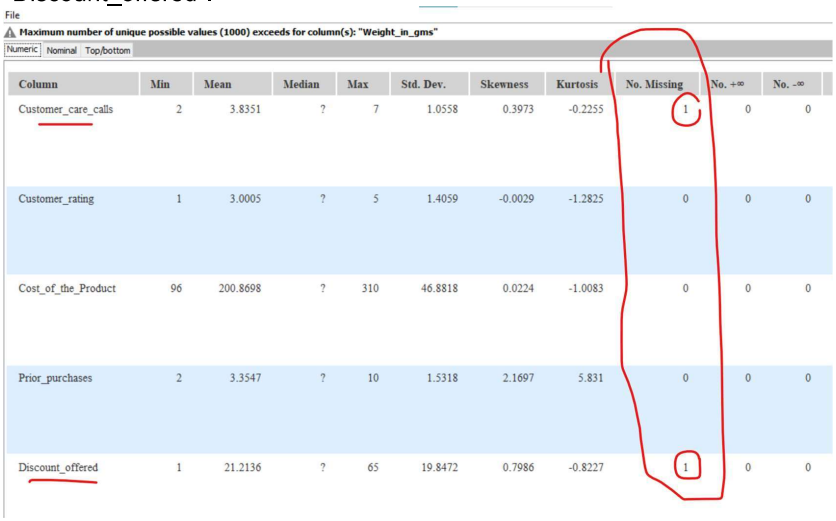
Commented [LSL7]: Provide some background information.
Refer to the question description.
Include some other research that are related to the topic.
Could be why the data was collected?
Or how the data was collected?
Or what can we learn from the data?
Or any past analysis done before?

Include your research references in section 10 of the report.

Give an overview of the data structure.

4. Data Cleaning

From the Statistics page, there was missing values in “Customer_care_calls” and “Discount_offered”.



File

⚠ Maximum number of unique possible values (1000) exceeds for column(s): "Weight_in_gms"

Numeric | Nominal | Top/bottom

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞
Customer_care_calls	2	3.8351	?	7	1.0558	0.3973	-0.2255	1	0	0
Customer_rating	1	3.0005	?	5	1.4059	-0.0029	-1.2825	0	0	0
Cost_of_the_Product	96	200.8698	?	310	46.8818	0.0224	-1.0083	0	0	0
Prior_purchases	2	3.3547	?	10	1.5318	2.1697	5.831	0	0	0
Discount_offered	1	21.2136	?	65	19.8472	0.7986	-0.8227	1	0	0

Continue …

Commented [LSL8]: Describe how do you know if there are any missing values in your data. What method did you use to resolve it? Show screen shot of your work flow and result.

If you are given more than one data set, you will need to either append them or merge them into one single file.
Explain how you do it.

If there are wrong data type, show how you convert it or resolve it.

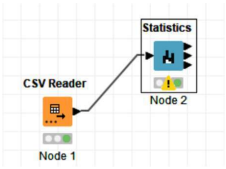
5. Exploratory Data Analysis

Commented [LSL9]: Perform statistical analysis of all the numerical variables.
Perform data transformation if needed.
Perform data mining to gain further insights. For example, plot a pie-chart or histogram to find out more.
Show and explain box plots (if any)
Show contingency table (if any)
Describe the percentage of probability.
Describe your findings.

Each sub-section can be analysis of each variable.
You can choose to analyse all variables or pick a few important ones to discuss (due to time constraint)

5.1 Statistical Result

Statistics result

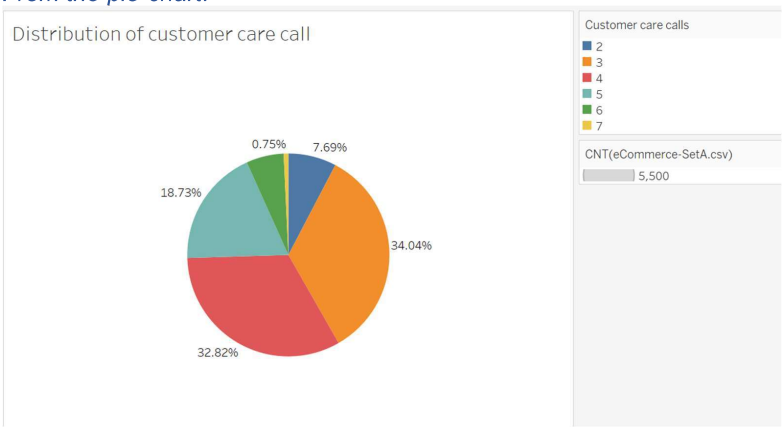


Variable	*Min*	*Max*	*Mean*	*Standard Deviation*	*Median*
Customer_care_call	*2*	*7*	*3.8351*	*1.0558*	*4.0*
Customer_rating	*1*	*5*	*3.0005*	*1.4059*	*3.0*
Cost_of_the_product	*96*	*310*	*200.8698*	*46.8818*	*199*
Prior_purchases	*2*	*10*	*3.3547*	*1.5318*	*3*
Discount_offered	*1*	*65*	*21.2136*	*19.8472*	*10*
Weight_in_gms	*1001*	*7846*	*3361.7889*	*1533.7764*	*3375*
Reached_on_time	*0*	*1*	*3.8351*	*0.4285*	*1*

5.2 Customer Care call

50% of the customer_care_call is less than 4, and Mean value is 3.8.
"Customer_care_call" should be interpreted as categorical data type to have a meaningful insight.
Customer calls at least 2 times and up to 7 times.
What is the mode?

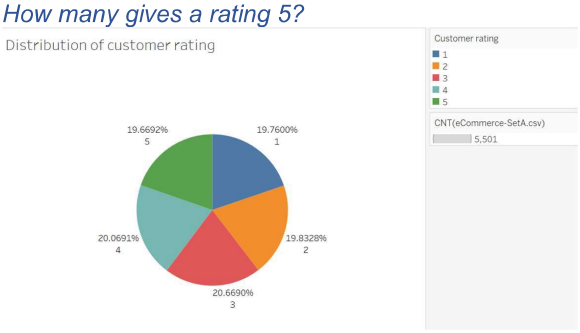
From the pie-chart:



Number of customers calling 3 times has the highest occurrences.

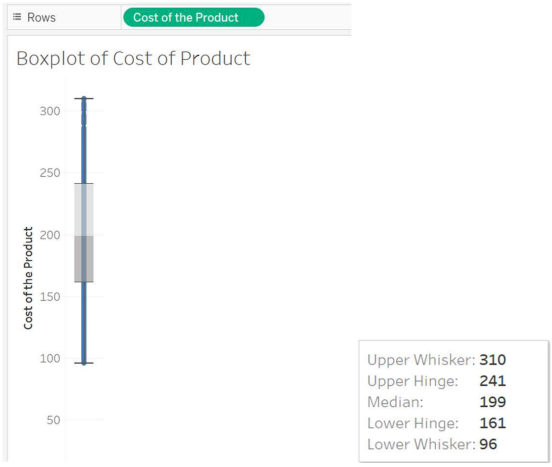
5.3 Customer Rating

Customer_rating should also be a categorical data type. There are 5 ratings. 50% give rating less than 3. It shows customer is generally satisfied.
How many gives a rating 5?



There are almost equal number of votes for each rating. Rating 3 and 4 are slightly higher.
:
:
Etc..

5.4 Cost of Product

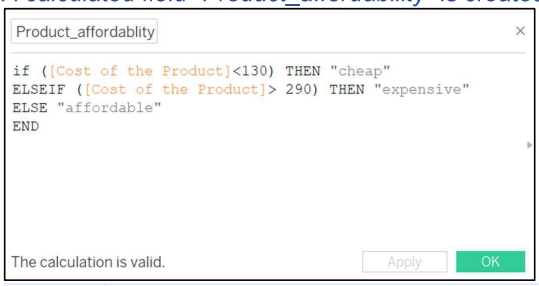


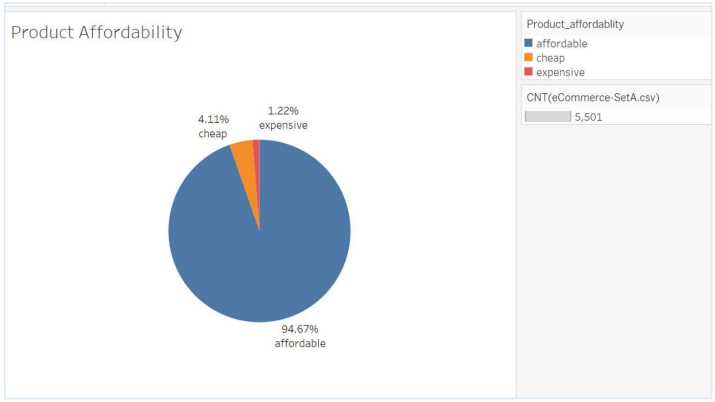
The box plot shows a good spread of values, ranging from 96 to 310. 50% of the product are lower than 199.

The histogram shows mode is between 150 and 170.
Product cost > 280 and <130 are at much lower number.

A calculated field "Product_affordability" is created for further insights:

Commented [LSL10]: This is data transformation. Seek opportunity to demonstrate data transformation skill in re-shaping your data for meaningful insights.

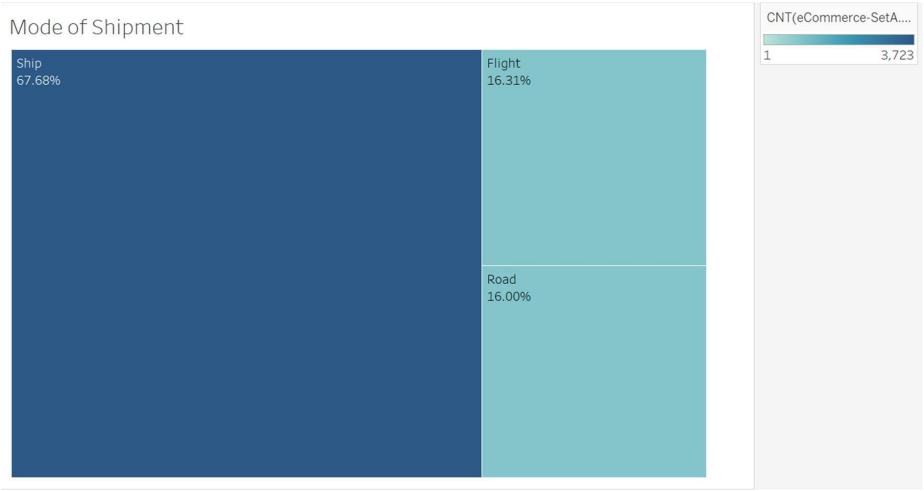




The pie-cart shows that product cost that are > 290, contribute 1.22%. Product cost between 130 and 290 are popular among the customers.
Continue …

6. Further Insights

Question 1: Which mode of shipment is popular?



Delivery by ship is most popular. There are almost equal number for Flight and Road mode.

Continue…

Commented [LSL11]: Further analysis with different visualization charts.
Answer your preliminary questions, if they were not answered in section 5.

Include all or some of the following (due to time constraint and depend on the data types):
Explore on Linear correlation, with correlation matrix and scatter plot (if any). Explain your insights.

Explore on contingency table (xTab). Explain the probability base on the given data.

Possible scenario of conditional box plot (numerical vs categorical)

Plot variable vs timeline (if any). By week, by day or by hour. Use appropriate aggregate method. Explain your insights.

Seek opportunity to perform data transformation for further insights.

Explore different visual charts to explain insights.

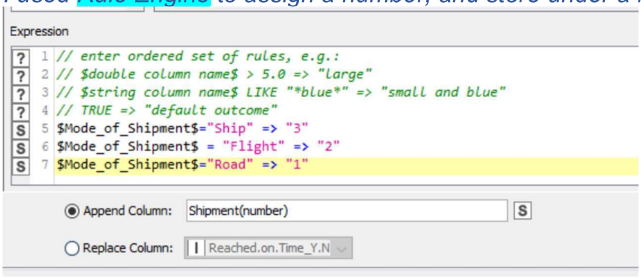
7. Data Modeling

Target for prediction is "reach on time". As the outcome is a categorical data type, logistic regression learner is used to generate the model.

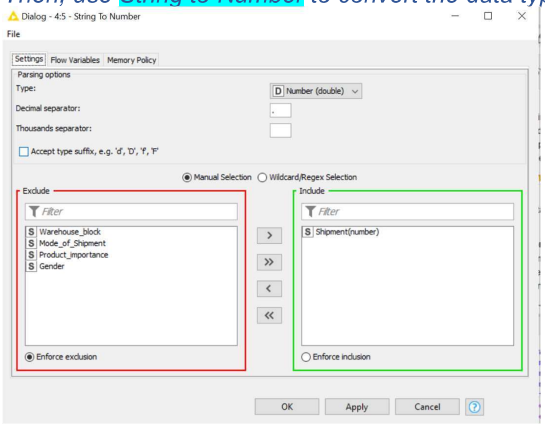
To use the categorical data type for training, they are converted to numerical using Category to Number node.

For mode of delivery, I will assign a 3 for shipment mode, 2 for flight and 1 for road. Because shipment will usually take longer time than road.

I used Rule Engine to assign a number, and store under a new variable.



Then, use String to Number to convert the data type.



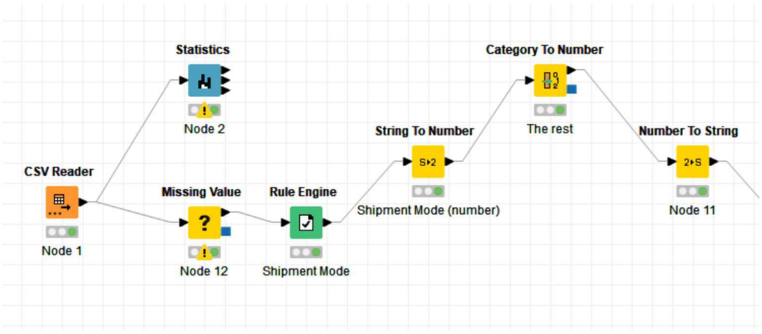
For the rest of the categorical data type, I use "Category to Number" to assign a random number.

Commented [LSL12]: Identify the target for your data.
It is either a linear regression or logistic regression model.

Check if you need to perform some data transformation and normalization, before continue with linear regression or logistic regression learner.

Explain your result.
What is the accuracy of the model?
Is the model suitable?
How can you improve the accuracy of the mode?

Refer to Student Guide, Pg111 for detail guides.



To predict "reach on time" as categorical outcome, it needs to be converted using number to String node, before continuing with data modelling.

Continue …

8. Conclusion

- There is inverse correlation between weight of the parcel and parcel reaching on time.
- The most frequent mode of shipment is by ship.
- Parcel delivered from warehouse B has the highest probability to reach on time.
- Etc.

Commented [LSL13]: Conclude your findings, insights.

9. Reflection

I was able to keep track of my progress to complete this project. I am now more competent in using KNIME and TABLEAU. I learned how to gather insights from data through data visualization. I choose appropriate visual chart base on the data type and the purpose of the variable. Data visualization and analytics skill can be further apply …continue …

Commented [LSL14]: Reflect on the process.

What are the challenges?
How did you resolve them?
What would you do differently?
What have you learned?
How can you apply the skills learned in this project in future career?

10. References

1. Ecommerce Guide (2021). Retrieved from https://ecommerceguide.com/guides/what-is-ecommerce/
2. Kaggle(2021). Retrieved from https://www.kaggle.com/prachi13/customer-analytics