Leitfaden für nachvollziehbare Schritte

1.Kurze Darstellung des Problembereichs / Aufriss des Themas

1.1Inhaltlich

Kern der Untersuchung

Grobziele der Arbeit : For any buisness its essential to have objectives in their plans well before the start of a new fiscal yeras. These companies can best reach thier goals by staying customer focused, offering products and catering their services to customers need.The real time data allows them to forecaste the potential sales and demand of thier items through predictive analytics. Marketing plans helps them to better define their target customers and store concepts.

Hence the goal of this analysis is to use the data and estimate the sales of the supermarket.

1.2 Begründung desThemas

Darstellung der Relevanz des Themas?

Warum ist das Thema wichtig und interessant und daher bearbeitungs- und förderungswürdig?

The goal of this analysis is :

1.See patterns in shopping.
2.Determine the right price that will attract more customers.
3.Implimenting campaigns and promotions.
4.Determining the demands of the specific product and targeting the audience.
5.Building Customer Satisfaction.
6. Increasing the store rating.
7. Increasing the revenue

Darstellung eines persönlichen Erkenntnisinteresses.

Dieser Abschnitt soll ein prägnanter Einstieg in die Projektarbeit / Seminararbeit sein.

Er soll beim Leser Interesse für das Thema und die Bereitschaft wecken oder verstärken, die Arbeit zu betreuen bzw. zu fördern und dient der Eigenmotivation.

The growth of supermarkets in most populated cities are increasing and market competitions are also high. Supermarket owners face major competition from other retailers and it very important to bring stability and growth in business to sustain in the market.Also the companies which constantly meet the needs of customers are typically more successful . Analytics empowers supermarkets to potentially increase the profits and improve the experience of each customer.

2. Nachvollziehbare Schritte

2.1 Der Stand der Forschung / Auswertung der vorhandenen Literatur / Tutorials ...

Wurde das Problem früher bereits untersucht?

Welche Aspekte wurden untersucht und welche nicht?

Welche Kontroversen gab es und welche Methoden standen bis jetzt im Vordergrund?

Yes the problem has been investigated pervious with applying different algorithm of machine Learning.

The chanllenges involved in the analysis is that the dataset have no attributes of selling price and the costprice also the exact location of the store is missing . which would affect the buisness in terms of increase in profit.

Lösungswege strukturieren!

Wichtigste (verwendete) wissenschaftliche Positionen zum ausgewählten Thema?

(Z.B. **Tutorials …**)

The analysis is done using BI tools Tableau.And data preprocessing and model building is done on KNIME Platform. Loaded the data set using CSV reader node and explore the data using Exatract table dimensiion node. Then the preprocessing is done using the column filter node. The data had some categorical columns which were converted to numeric using the category to number node.

2.2 Fragestellung:

Can the analysis of the data increase the profit ofthe buisness and customer traffic

2.3 Wissenslücke:

The data set had no selling price and cost price attibute . So its difficult to analyse wether the transaction made profit or loss . Also the location of the store plays important role in customer traffic in the branches . There was no information of the other retail stores near the branches making it difficult to analyse the potential compitators

2.4 Methode

Detaillierte nachvollziehbare Beschreibung der Vorgehensweise !!

Vgl. MUSTER-PROJEKTE in den Tutorials !!

Loaded the data set in tableau and KNIME

Analsis is done using bar graph , scatterplots and bubble plots.

Data preprocessing is done in KNIME using Nodes avalaible .

Simple linear regression model, Decession Tree and Random Forest is used to buid a model

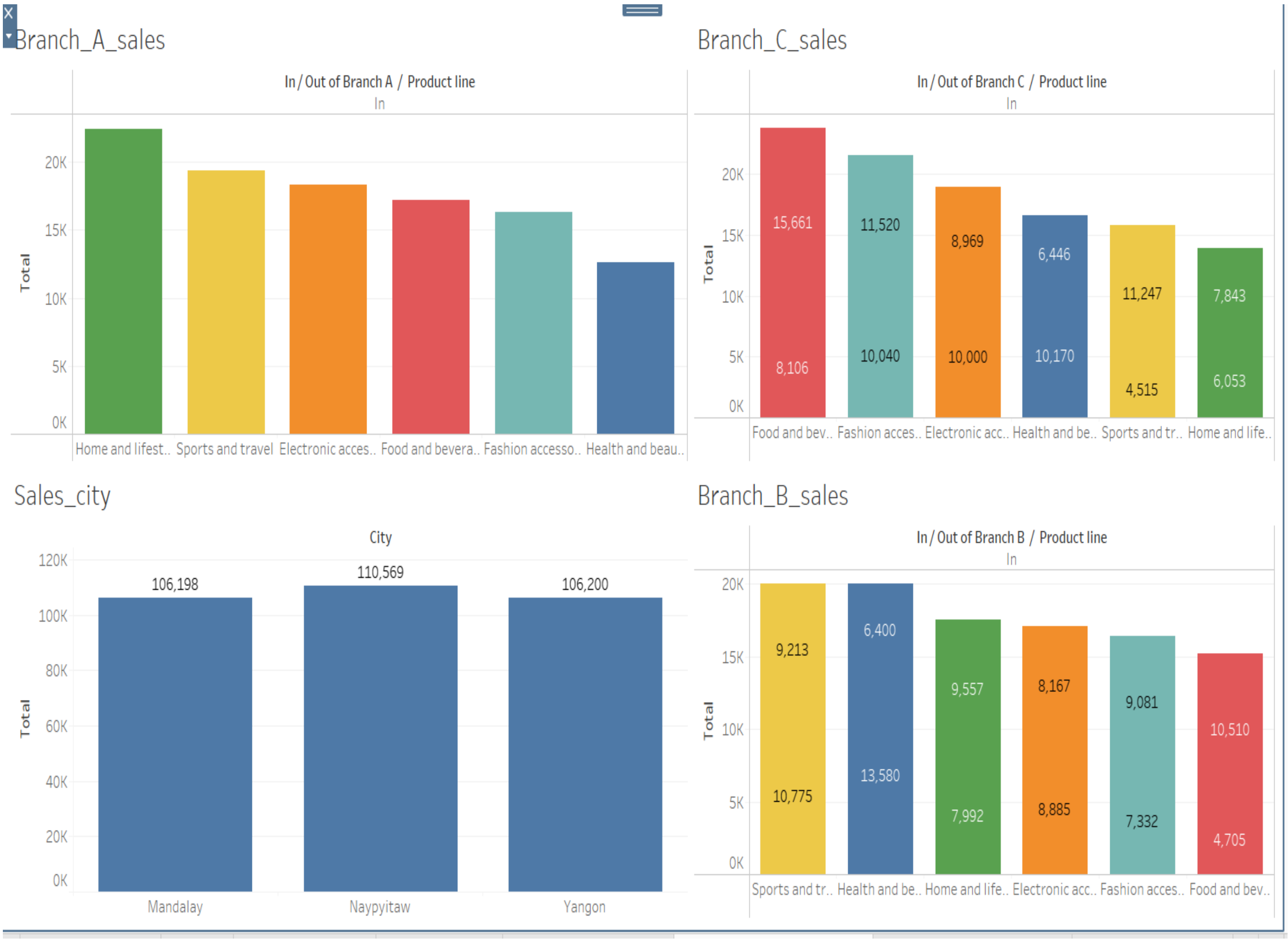
Analysis 1

1. What was the total number of sales? Which city has the highest number of sales?

The city Naypyitaw seems to have more buisness than other cities. Mandalay and Yangon almost make same sales

2.Which branches have more sales from which product category?

Branch A have more sales on Home and Lifestyles . Branch B has highest sales on Sports and travel and health and beauty. Branch C makes more profit on Food and Beverages.



Analysis 2

1. What type of product is sold the most?

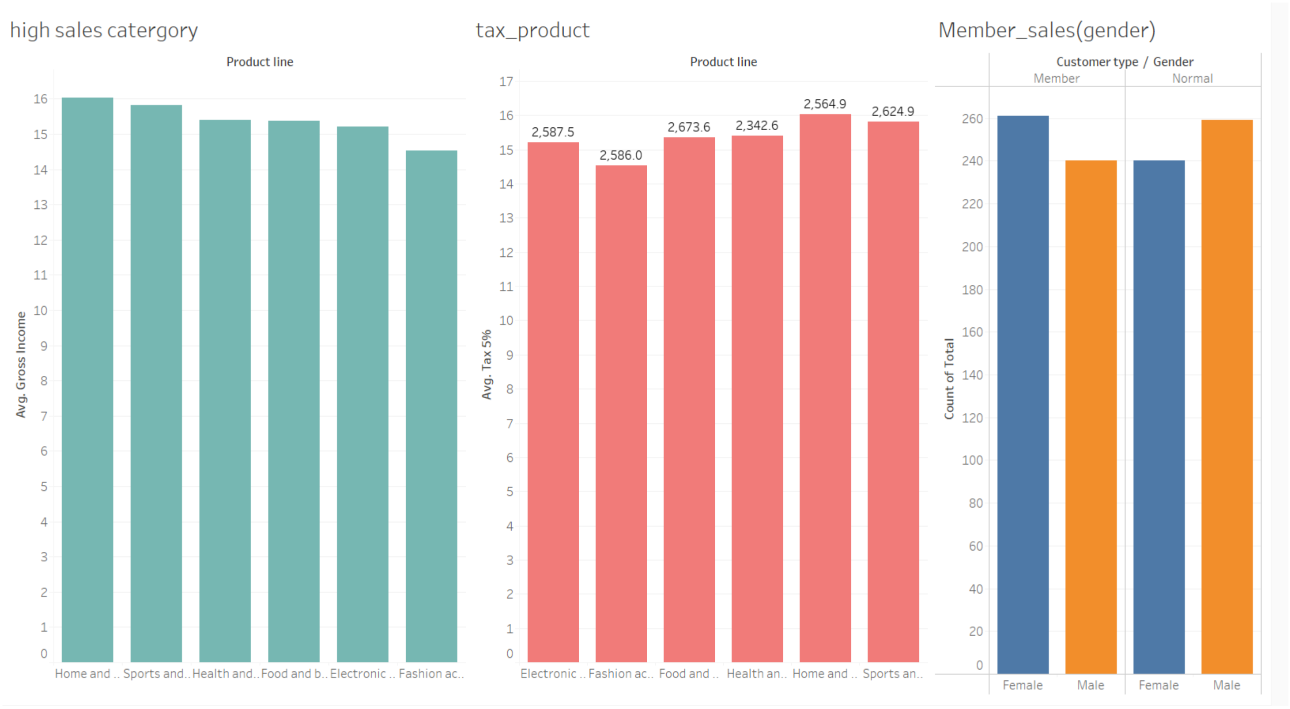
The Home and Life style products are sold more . And the second highest sold product is Sports and Travel. Also the supermarket sale is pretty less in the category Fashion Accessories. So it needs a marketing campaign and more advertisments

2.Which Category product have high average tax?

On an average tax colloected from each category are the same . But home and Life style and Sports acccessories seems to do pretty good interms tax.

3. whats the spending pattern of Female and Male ? Are they memebers or Normal customers?

The spending pattern doesn't have much difference in Female and Male . But Female custmors shop more who are members and the non members customer males are the bit higher



Analysis 3

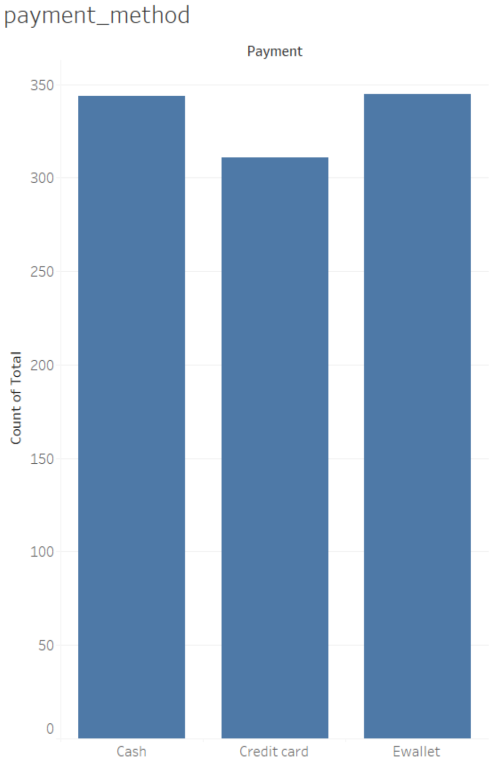
1.What is the favorite method of payment of the customers?

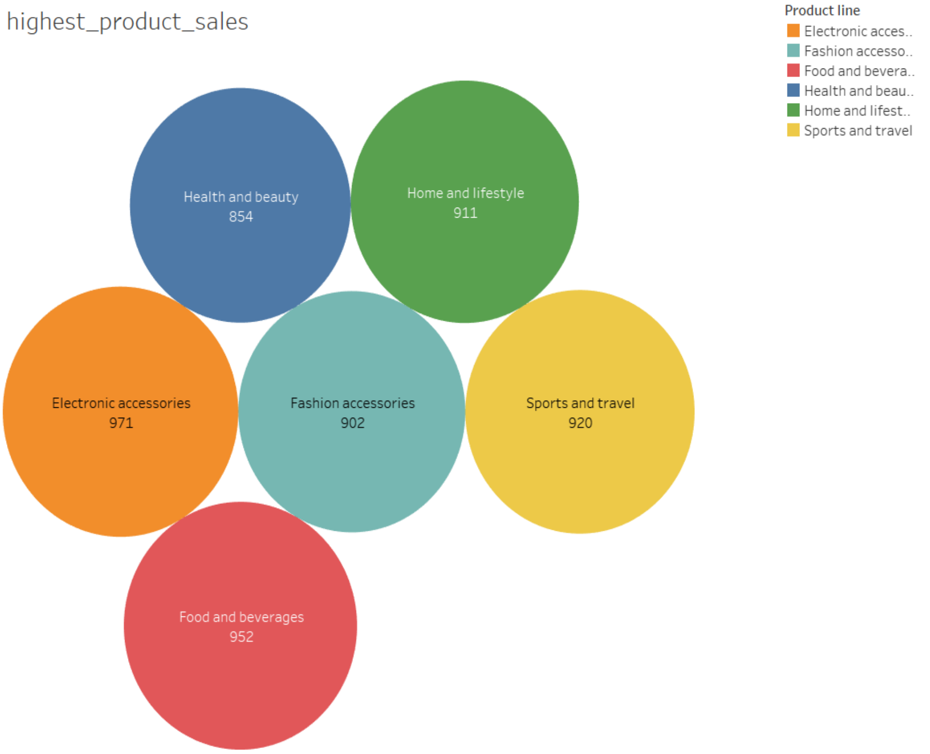
The most popular payment method is in-fact E-wallet and not credit cards.

Cash payment is also popular.

2. which are the products with high number of sales ?

Electronic Accessories seems to have more sales compared to other but in general all other categories also have eqaully good number of sale . Health and beauty products seems to have less Count of sales .





Analysis 4

1.Which gender spend more?

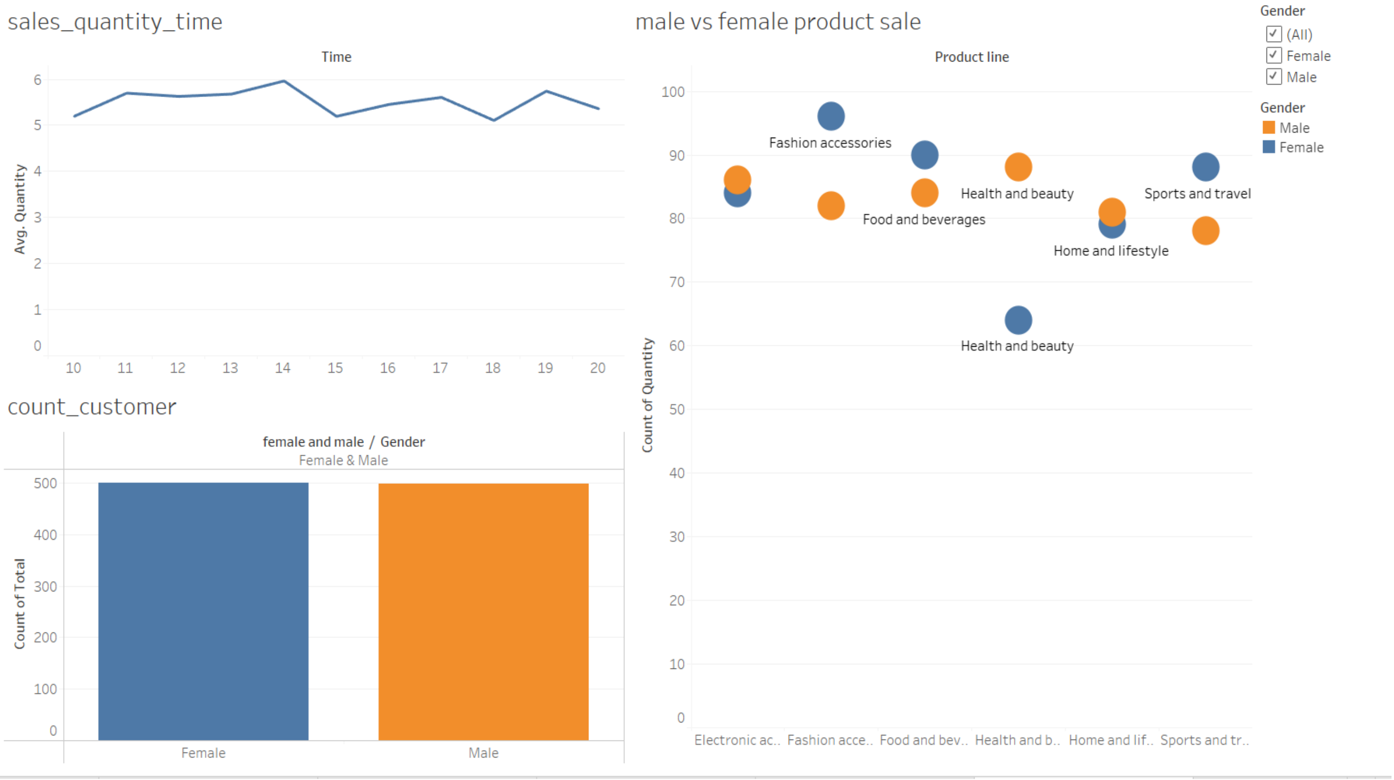
The graph clearly shows that there is no bias with gender men and women spend equally.

2.What time should we display an advertisement to maximize the revenue?

Peak is observed in the 14th hour i.e 2 pm of the day. Hence, sales is typically higher in the afternoons.

3.What gender buy more items in each category? what is the category?

Females spend on 'fashion accessories' the most and for males it is 'Health and beauty'. Females also spend more on 'Sports and travel'.Female and male spend equally on electronics.



Analysis 5

1.Which day of the week has maximum sales?

Sales is highest on Saturdays probably because it is the weekend. Interestingly,Tuesdays is a close second.Mondays is the lowest in sales, probably because it is start of the working week.

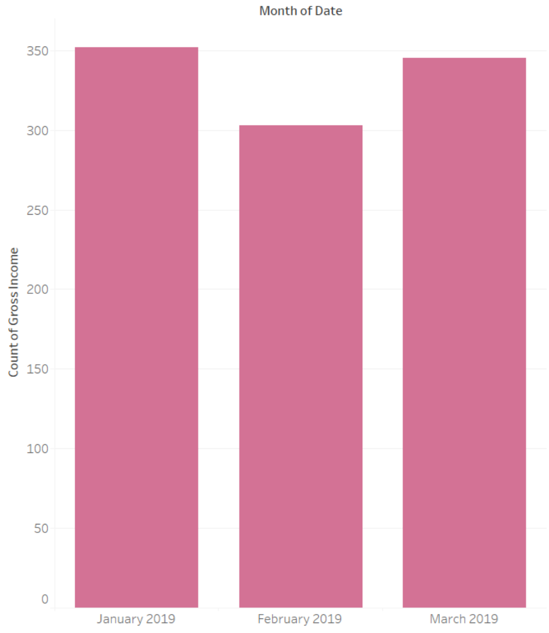
2.Which month have more sales?

From the graphs its clear that january month had slightly more sales compared to February and march.

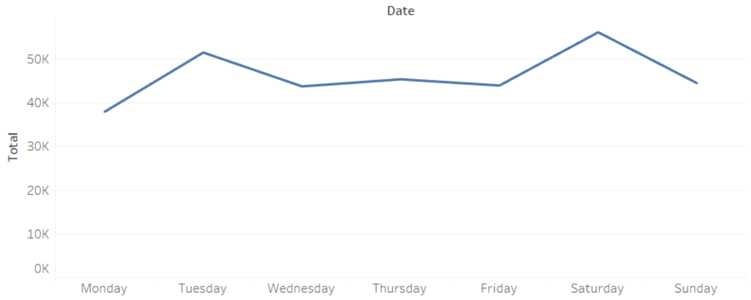
3. Which branch had highest rating ?

Looking at the graph there is no much difference in all three branch. Branch A has slightly more ratings than other branches

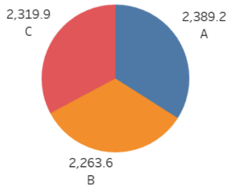
Grossincome_month



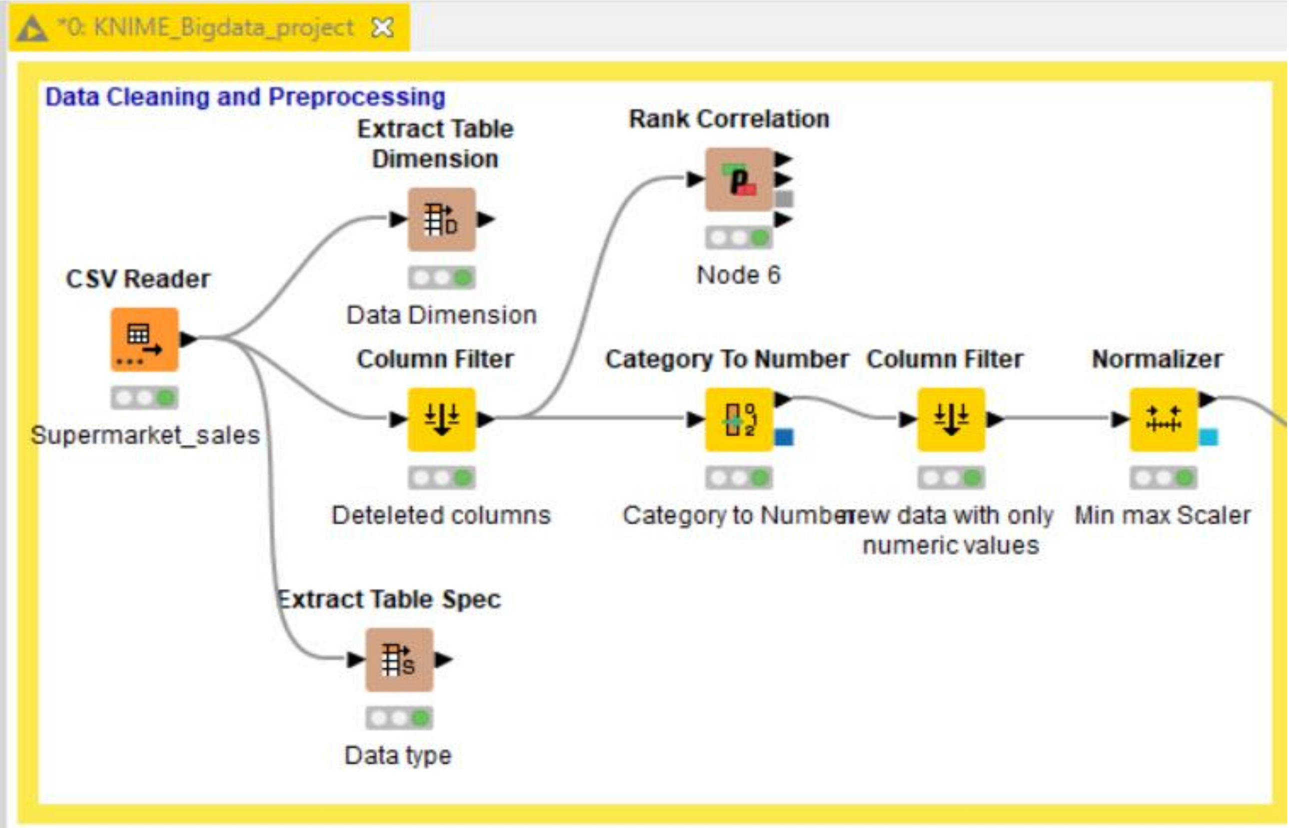
Sales on weekdays

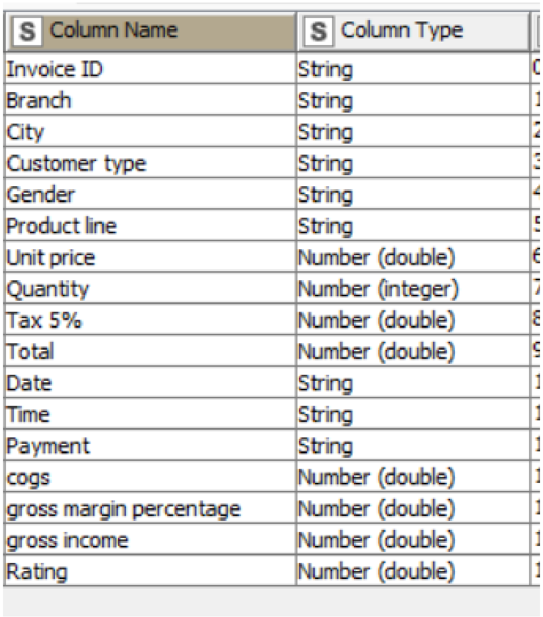


Rating_branch



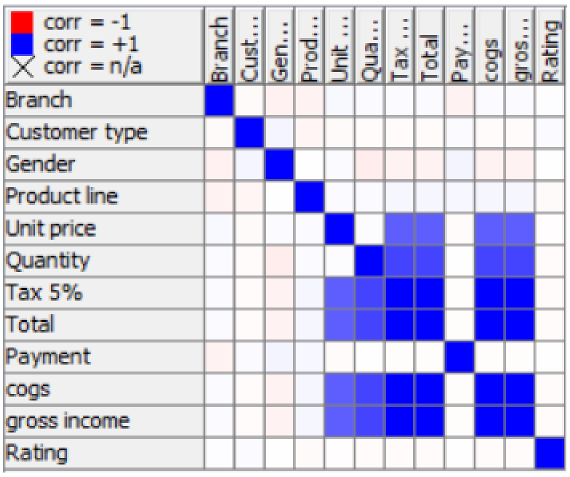
Data Preprocessing and Machine Learning:





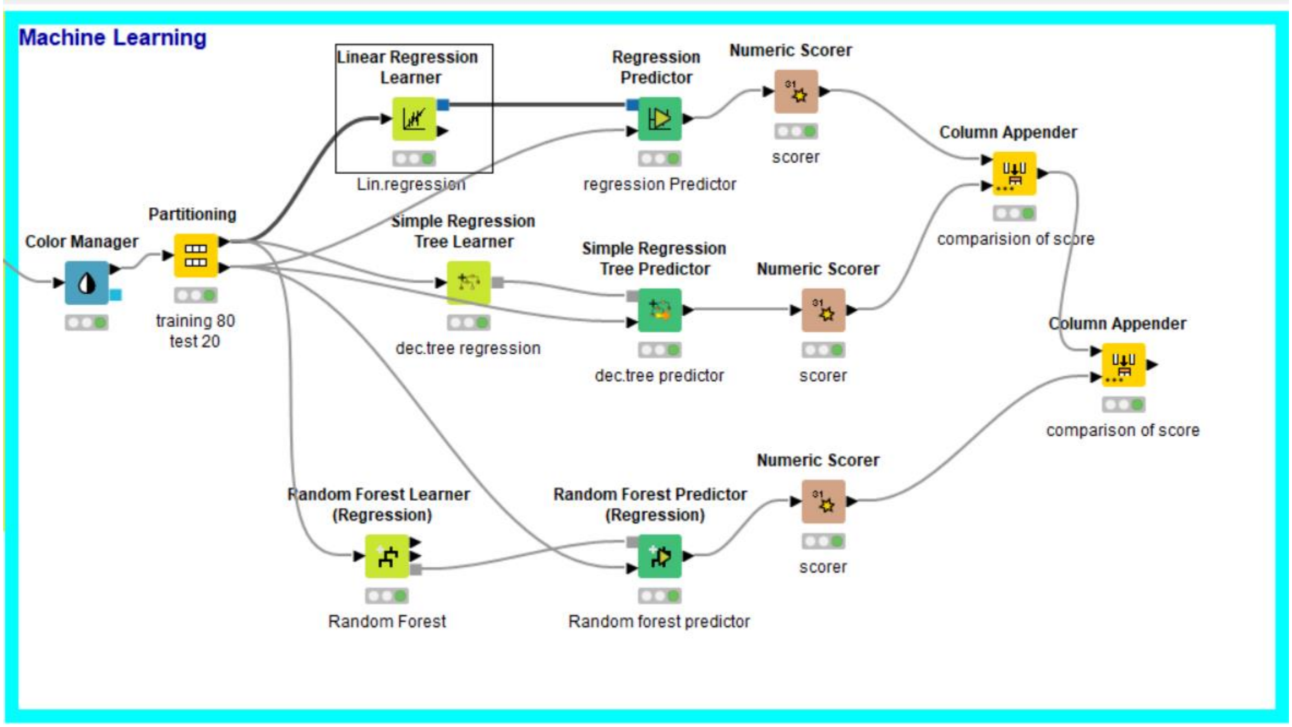
Column Name	Column Type
Invoice ID	String
Branch	String
City	String
Customer type	String
Gender	String
Product line	String
Unit price	Number (double)
Quantity	Number (integer)
Tax 5%	Number (double)
Total	Number (double)
Date	String
Time	String
Payment	String
cogs	Number (double)
gross margin percentage	Number (double)
gross income	Number (double)
Rating	Number (double)

I have used the CSV file reader to read the table then with the help of Exract table dimension explored the data.Also i have used the column filter node for deleting the column which we dont need for model buiding. The data set contains 1000 instances and 17 attributes. There are arround 9 columns which contain Catogorical column which has been converted to numerict using the node Catogory to numeric.After coverting i have deleted the old columns with column filter node. To build the regression model i have used min max scaler for normalizing the data.

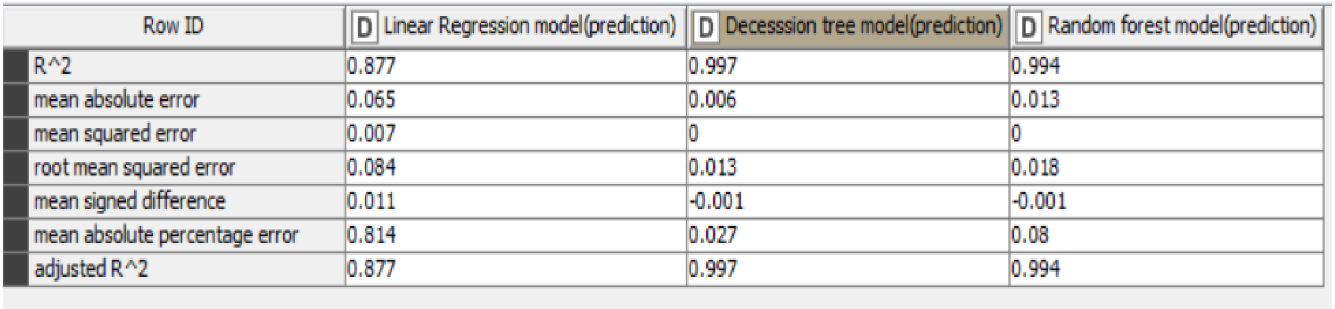


The corelation matrix clearly sho the some strong corelation beteween tax, quantity , unit price payment and cost ofgoods sold.

Now its time to move to Machine Learning part .The data set is partitioned in 80:20 ratio for training and testing process.I have used three methods the first one being Linear Regression Learner and the second one is Simple Regression Tree Learner and the last one Random Forest Learner .



2.5Ergebnisse



Row ID	D Linear Regression model(prediction)	D Decesssion tree model(prediction)	D Random forest model(prediction)
R^2	0.877	0.997	0.994
mean absolute error	0.065	0.006	0.013
mean squared error	0.007	0	0
root mean squared error	0.084	0.013	0.018
mean signed difference	0.011	-0.001	-0.001
mean absolute percentage error	0.814	0.027	0.08
adjusted R^2	0.877	0.997	0.994

2.6 Ausblick

The aim of this project was to analyse the supermarket data set and investigate which machine learning model yields the best performance .Utilizing the data provided the forecaste were implemented for all the three stores and products to analyse the sales quanity of each product category over a period of three months. From the above its very clear that the Decession Tree has the high accuracy of 99.7%