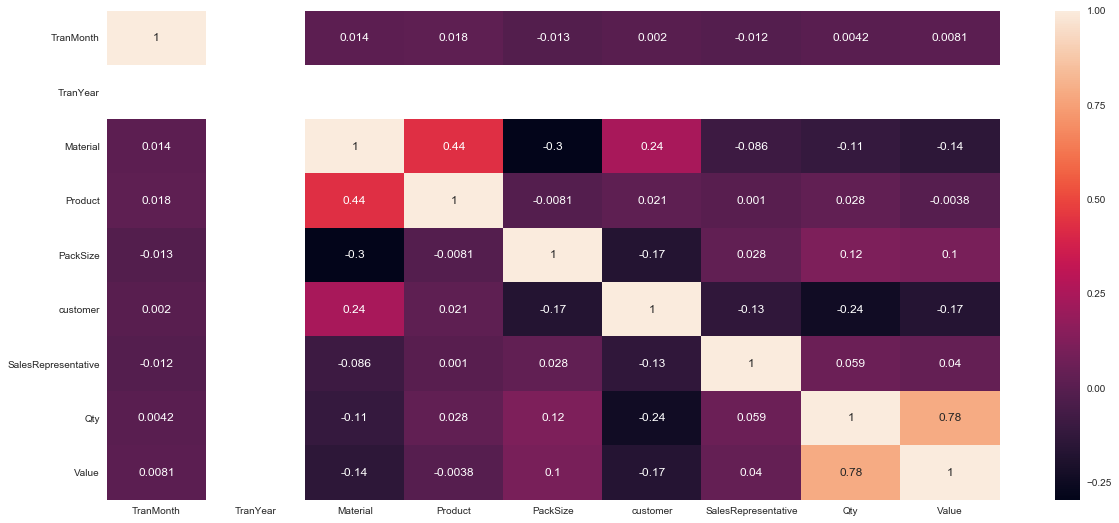
**Sales Data Correlation matrix:**

The starting point of **Exploratory Data Analytics** is a **correlation matrix**, in which the intercorrelations between the studied variables are presented.

Hence, first of all I prepared a correlation matrix for Sales Data provided.



As we can see here, Data is telling us that:

1. **Material** and **Product** columns are having correlation of 0.44
2. **Material** and **Customer** columns are having correlation factor of 0.24
3. **Quantity** and **Value** columns are having correlation factor of 0.78

And, as per point 1 Material-Product and Point 2 Material-Customer,

1. **Customer** and **Product** relation should be considered.

This gave me answer regarding which columns I should focus on to prepare charts.

And, what charts (with specific column combinations) I should prepare out of many possible charts that can be prepared.

**Important Columns**: 1) Material 2) Product 3) Customer 4) Quantity 5) Value

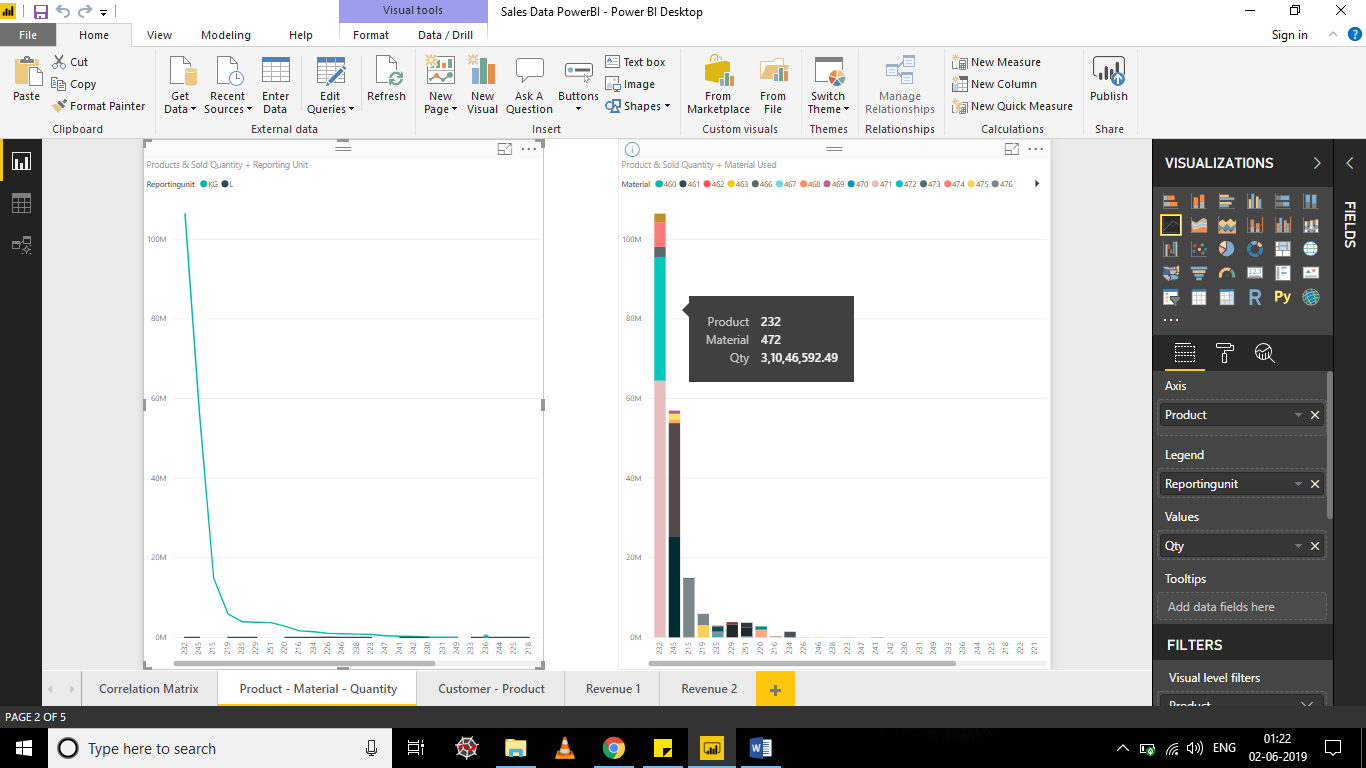
**#Product and Material relation**

As we are working in 2 dimensional plots, It is necessary to keep at least one axis univariate.

It is not possible to create a chart with Product and Material as both features contain multiple categories.

So, I prepared a chart of **Product VS Qty** and added **Material as Legend** to it.

**As the Tooltip suggests:** Material 472 is being used in Product 232 manufacturing. Qty used is approx. 31046592.



**From Client perspective, Insights we get here:**

**#1** for a particular Product, which raw materials are being used?

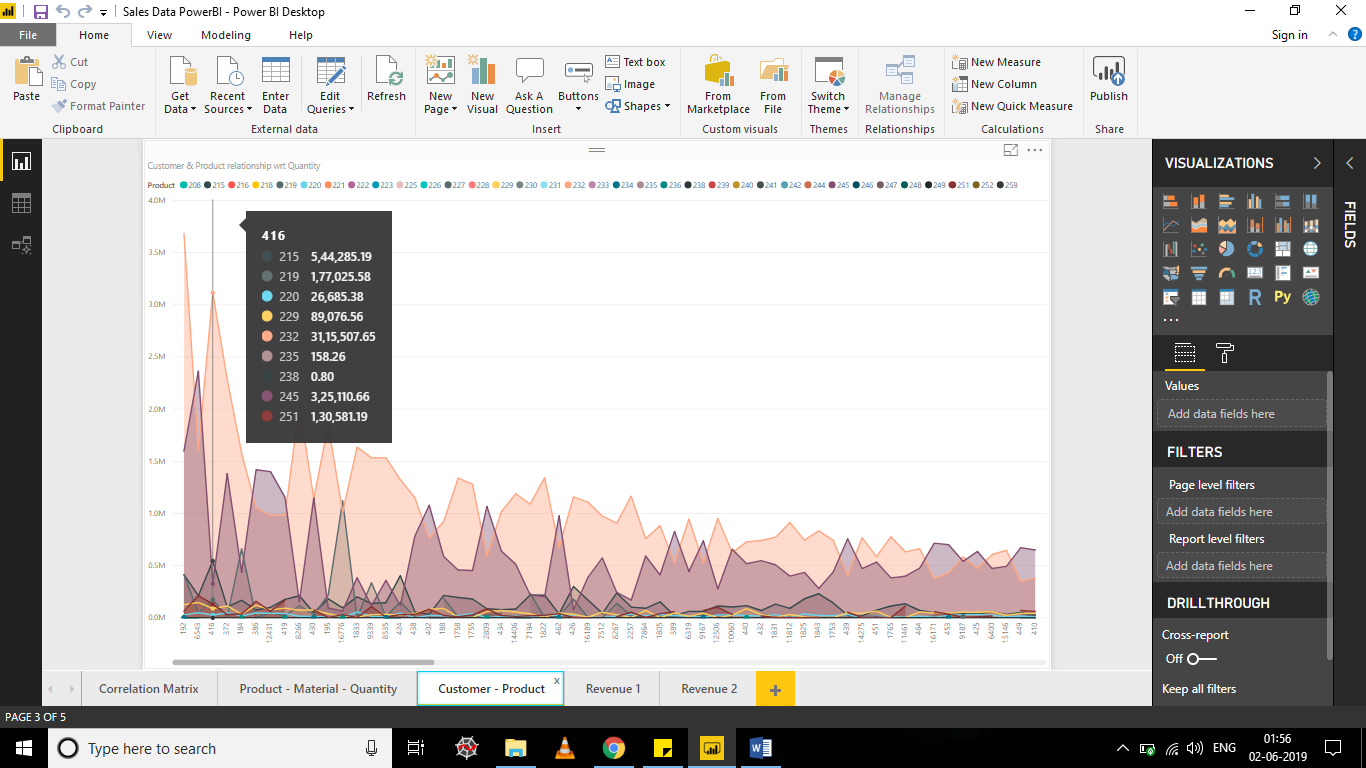
**#2** estimation of quantity of a particular material required in 2019?

**#3** At Whatever Production Branch and Production line, Product 232 is being manufactured, Materials used need to be made available there in an estimated quantity.

**#4** exporting of raw materials, Manpower & Equipment related arrangements, and Prepared Product Storage Planning could be estimated and executed efficiently.

**#Customer and Product Relationship**

A chart of **Customer** **VS Qty** and added **Product as Legend** to it.



**From Client perspective, Insights we get here:**

**#1** product order pattern of a customer? And, in what quantity orders are being placed by customer?

**#2** estimation of quantity of a particular product to be kept in nearby inventory – to avoid scenarios like,

“Sorry, product xyz is out of stock. We will deliver you in 2-3 working days.”

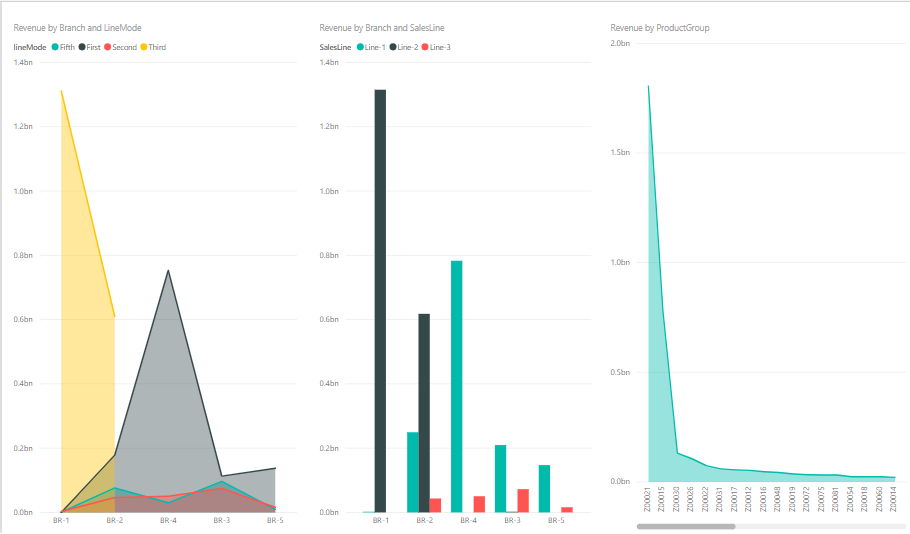
Out of Stock periods cause business lose to both dairy and shopkeepers.

**#3** Recommended products – example. If 5 rupees dairy milk is ordered, recommend 2 rupees Cadbury shots to customer during sales visit.

Better insights to Sales person – better recommendations – more orders – more business

**#4** availability of products to inventories near to customer locations and Product Delivery related logistic arrangements could be estimated and executed efficiently.

**#Value (Revenue)**



**Revenue by Branch and Production Line mode:**

**#1** Dairy management will get an idea about –

(1) Branch performance is as per expectation (2) which branch is under performing

(3) Why underperforming – equipment difference OR employee reskilling need OR poor execution of processes

**#2** if a branch and production line is performing as per expectations **–**

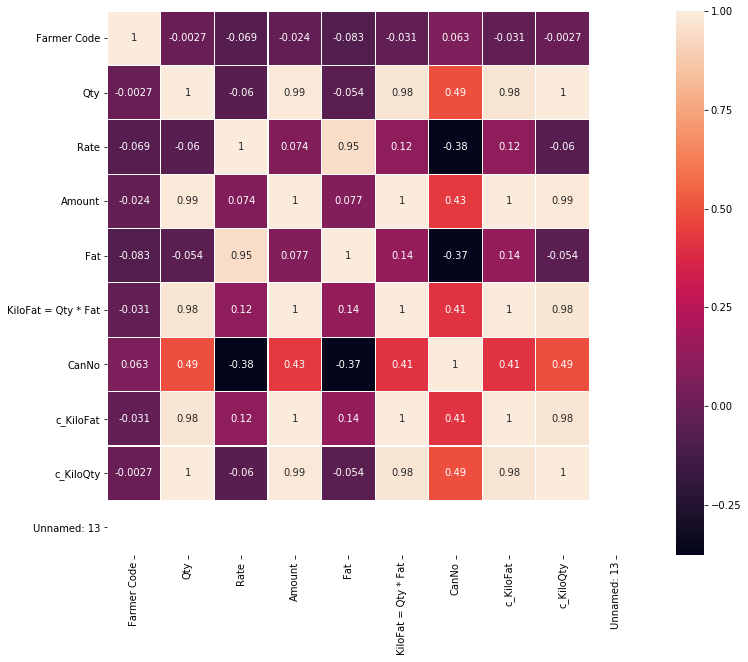
(1) Estimated Level of maintenance required and its frequency

(2) Reward branch workers and line workers appropriately for their hard work and dedication

**Revenue by Branch and Sales Line** & Other graphs prepared in **Revenue1** and **Revenue2** slides can be referred for:

**#** expected performance – Improvement required if any – better rewarding system – Peer to peer training requirement

**Milk Collection Data Correlation matrix:**



As we can see here, Milk Collection Data is poorly correlated. Reason is as follows:

1. **Amount** = Qty\*Rate **(2)** **KiloFat** = Qty\*Fat  **(3) c\_KiloFat** = KiloFat/100 = (Qty\*Fat)/100 **(4)** **c\_KiloQty** = Qty

All these four numeric columns are directly proportional to **“Qty”** column. (Refer Qty column in matrix)

**FarmerCode**, **Date** and **EntryDt** – columns do not provide any significant correlation with other columns.

**CanNo** – is of no significance

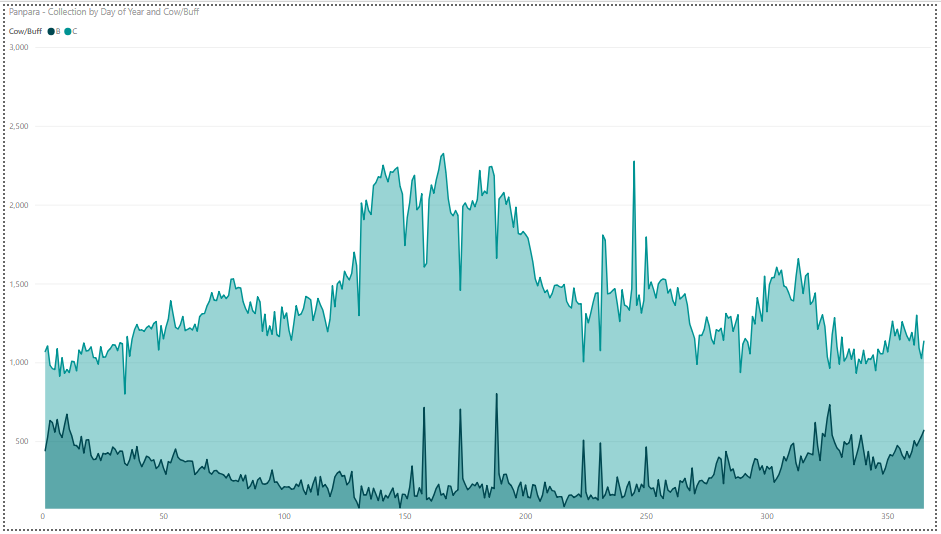
So, we have 5 (non-greyed column in below image) out of 13 columns to be focused on:

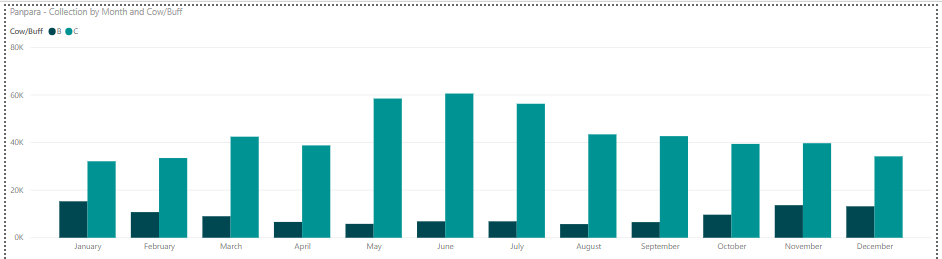
C:\Kunjan Data\Machine Learning\Spider Projects\PromptSoft Assignment\Milk Collection\Columns Relation.PNG

**So, Important Columns** are: 1) Shift 2) Cow/Buff 3) Qty 4) Rate 5) Fat

**#Qty**

I prepared a chart of Milk Collection from **Panpara - Qty VS Day of Year** and **Qty vs Month** and added **Cow/Buff** as legend to it.



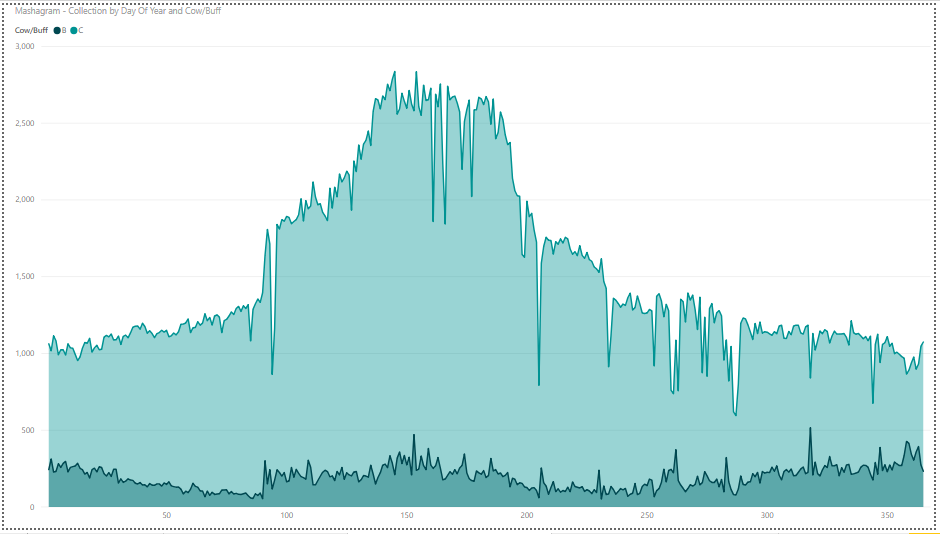


#1 during month of **May, June and July** – Cow milk collection is drastically increasing.

#2 as year progresses, we are having **gradual increase and then gradual decrease** in **Cow** milk collection.

And, **opposite** behaviour in **Buffalo** milk collection.

**Similar Behaviour** can be observed with **Mashagram Milk Collection Data** too. (Following image)



**#Insights**

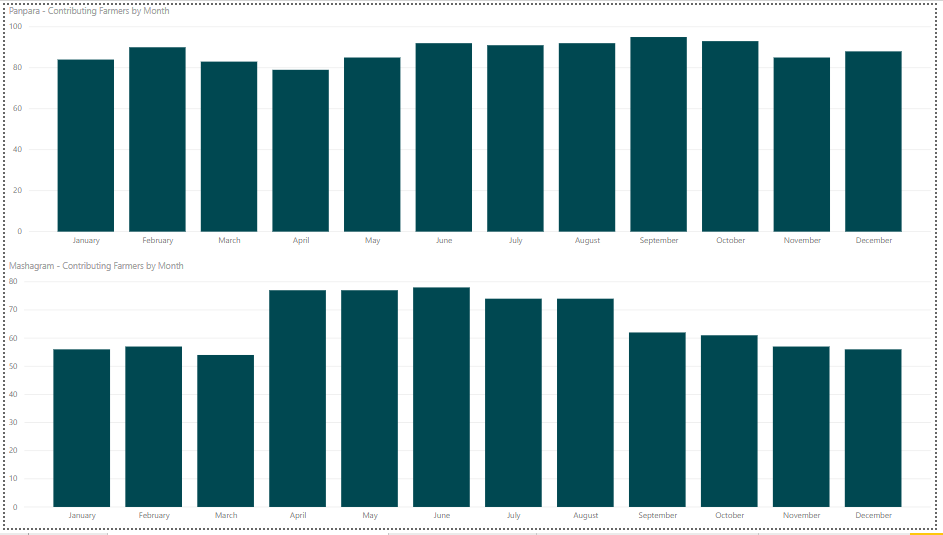
**#1** Quantity of Products manufactured from cow milk will be high during May, June and July

**#2** if sales of these products are not as per production capacity, other arrangements need to be done to utilize collected milk during these months.

**#3** Man power, Milk Collection, Cleaning and pasteurising, product manufacturing capacity, Storage facility and Sales support required will be high during these months

Question: why such pattern?

So, I plotted plot of **No. of Farmers depositing milk to Dairy VS month**. (Below image)



As we can see here, **Mashagram** data is showing **full support** to the claim that:

**No. of Farmers depositing milk to Dairy is increasing during April, May, June, July and August.**

Which leads to rise in milk quantity collected during these months.

But, **Panpara** data is giving **partial support** to my claim.

Which opens **possibility for some other factor** which is also driving this behaviour.

May be, its biological that cows produce more milk and buffaloes produces less milk during warmer days.

I am not sure. Reason need to be identified.

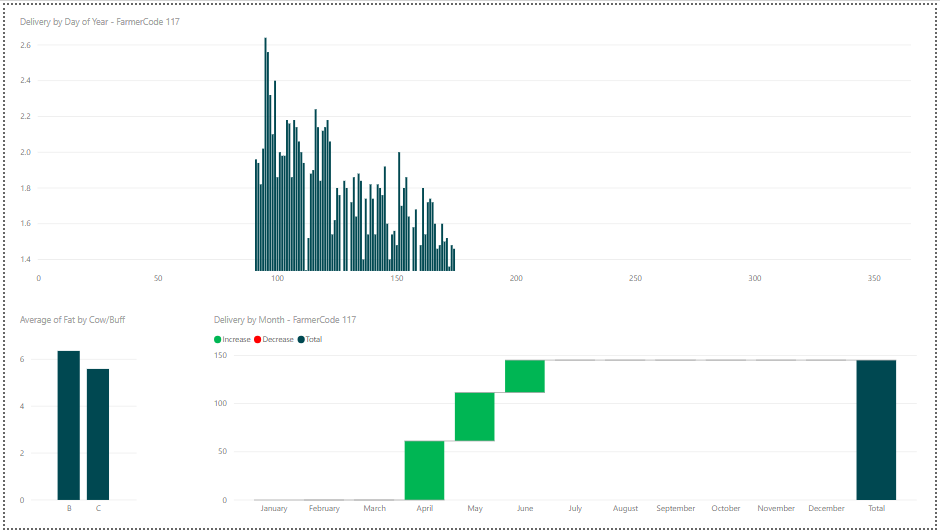
**Insight:**

**#1** why during particular months only some Farmers are depositing milk to dairy?

**#2** what can be done to attract those farmers to deposit milk throughout the year?

**#3** more milk – more production – more sales – more business

**One of such Farmer is FarmerCode 117:**



As we can see here, FarmerCode 117 is contributing during April, May and June.

And, as months progress Milk quantity deposited is decreasing gradually.

**Regarding Predictive Analysis of Milk Collection Data:**

"**Multicollinearity**" refers to predictors that are correlated with other predictors.  Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other.

It is a rule in predictive/regression analysis that if you have two or more features with a high correlation, remove one from the model. Because, it just cause calculation overhead to the model.

**Considering Qty feature and its relation with other features of data discussed above,** we are again down to 5 variables:

1) Shift 2) Cow/Buff 3) Qty 4) Rate 5) Fat

And, I didn’t feel confident about predicting **Qty** using **Shift, Cow/Buff, Rate** and **Fat** values.

Hence I have not prepared Predictive model of Milk Collection Data.

Thank You,

Kunjan