SYMBIOSIS CENTRE FOR INFORMATION TECHNOLOGY (SCIT)
Constituent of Symbiosis International (Deemed University), Pune
(Established under Section 3 of the UGC Act of 1956 vide notification number F-9-12/2001-U-3 of the Government of india)
Re-Accredited by NAAC with 'A' Grade

DATA VISUALIZATION AND MODELING

Assignment - 2

Case 02: Happy Cow

Prepared by:

DIVISION: D

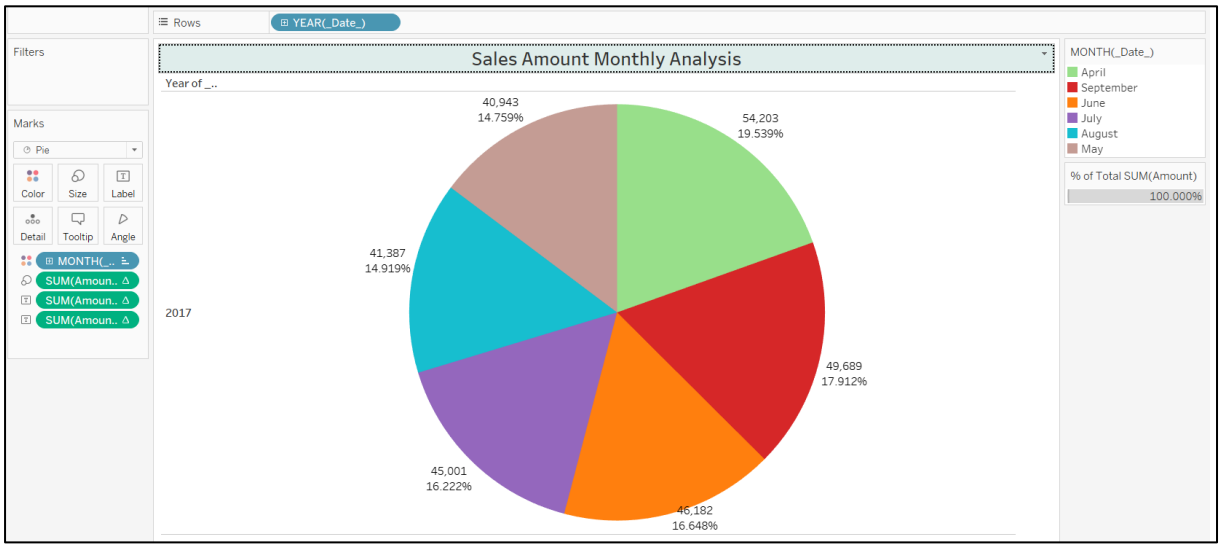
Name of Student	PRN NO
Suraj Kumar Shrivastava	20030241196
Date	4th Feb, 2022

Submitted to: **Prof. Krishna Singh**

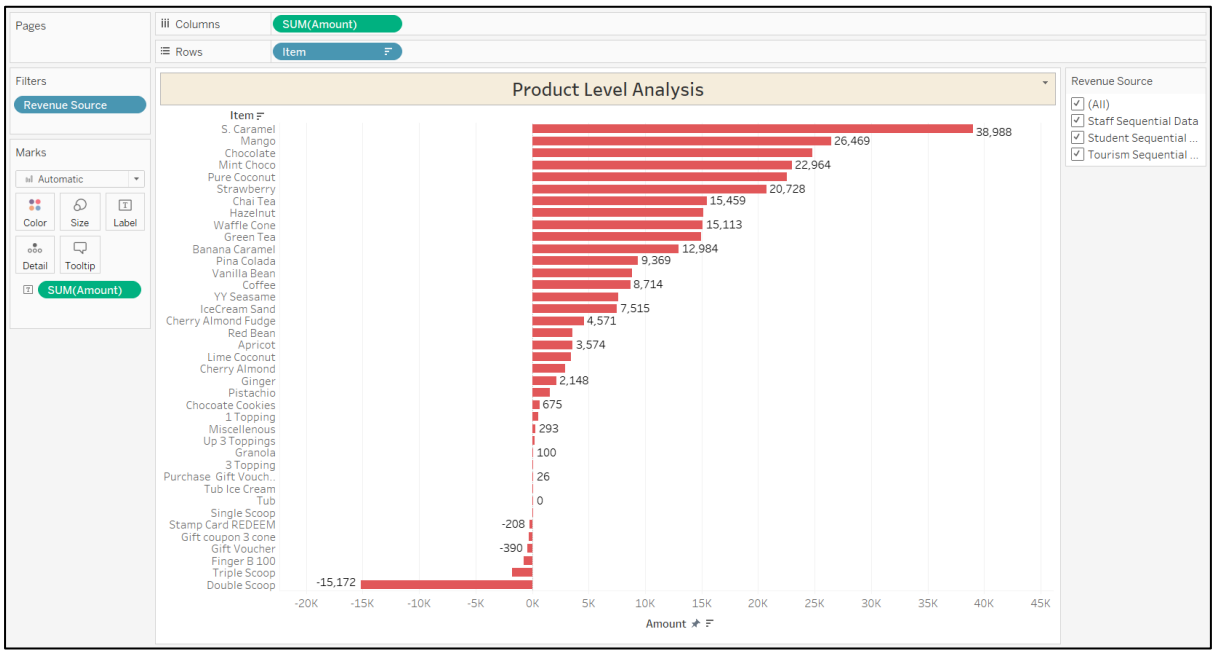
Analyzed Using tableau

Ques. 1. Explore data and forecasting products by using visual analytics.

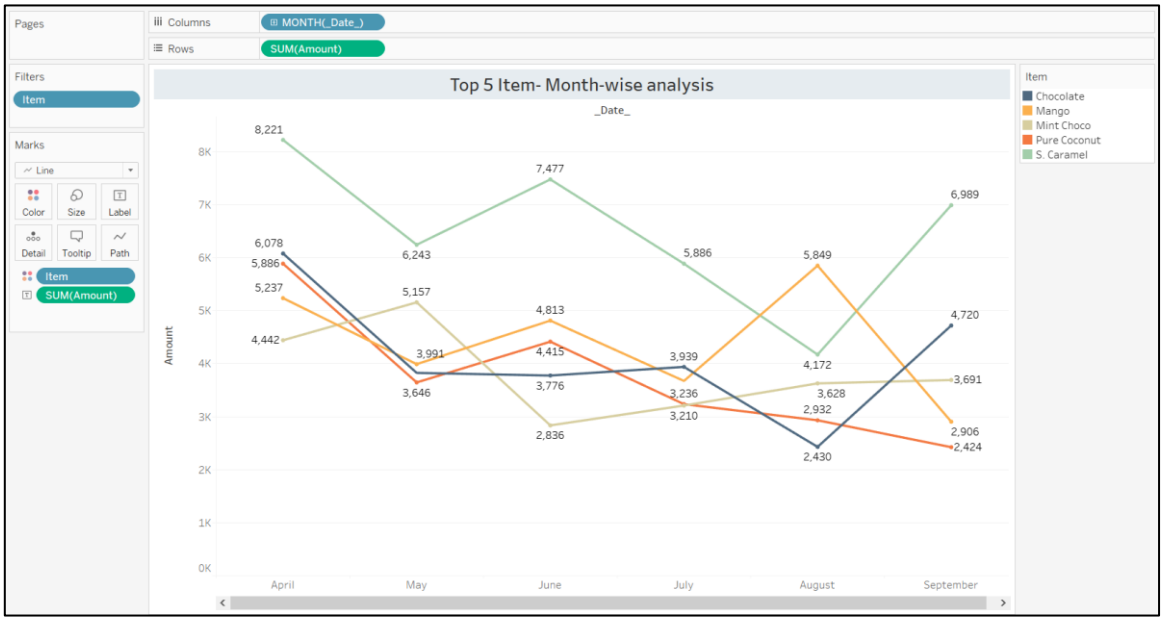
Ans. The data exploration part starts from the monthly analysis of sales amount, given below is the total amount & its percentage of each month:



Given below is the product level analysis of all the item sold, here we found the top selling products and also taken Student, Tourist and Staff as a filter which gives the liberty to visualize the datasets:

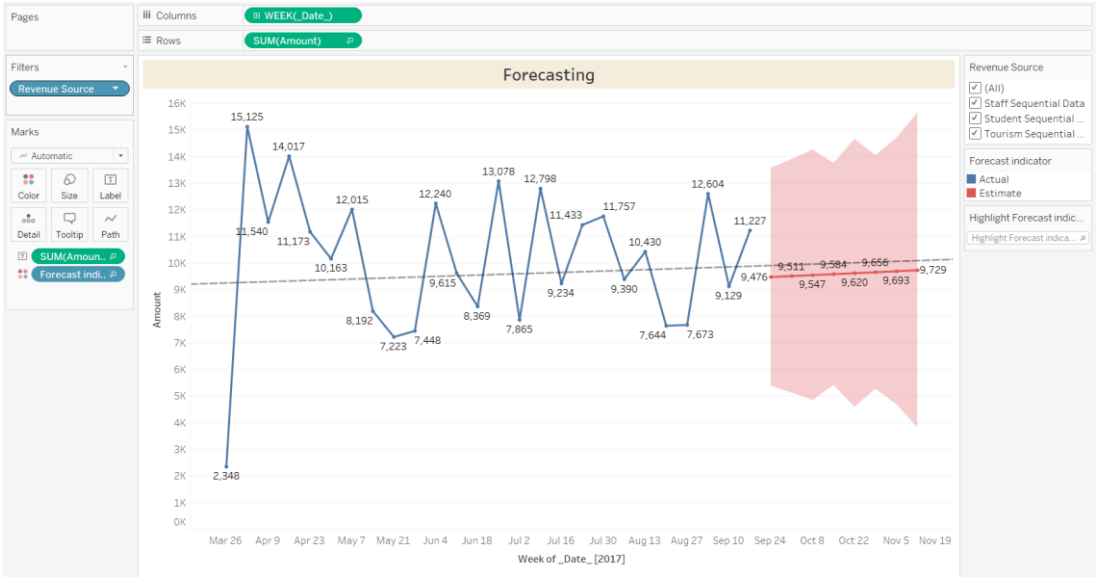


I have identified that the top-5 sold items are *S. Caramel*, *Mango*, *Chocolate*, *Mint Choco*, *and Pure Choco* and their monthly trend:



Forecasting:

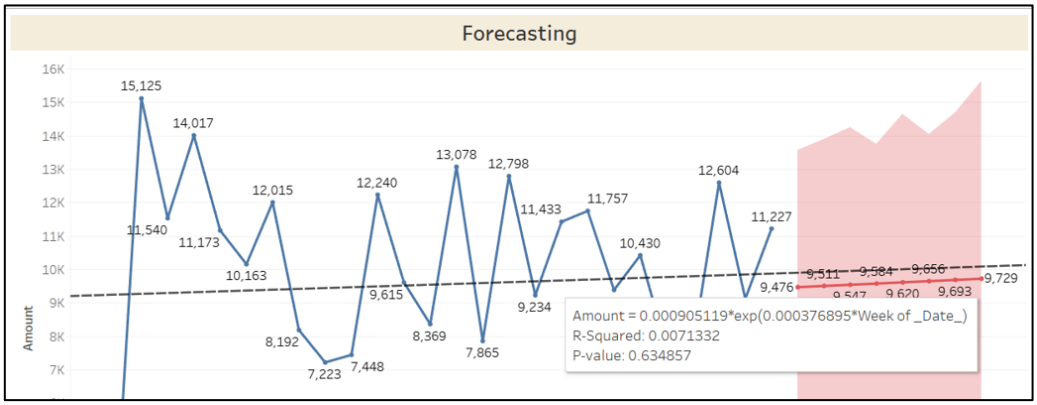
Forecasting the Sales values for Happy Cow dataset, I was able to estimate the sales values for next 8 weeks with prediction interval at 90% and trend as “Multiplicative”:



The accuracy of the prediction can be measured using the following values that we obtained:

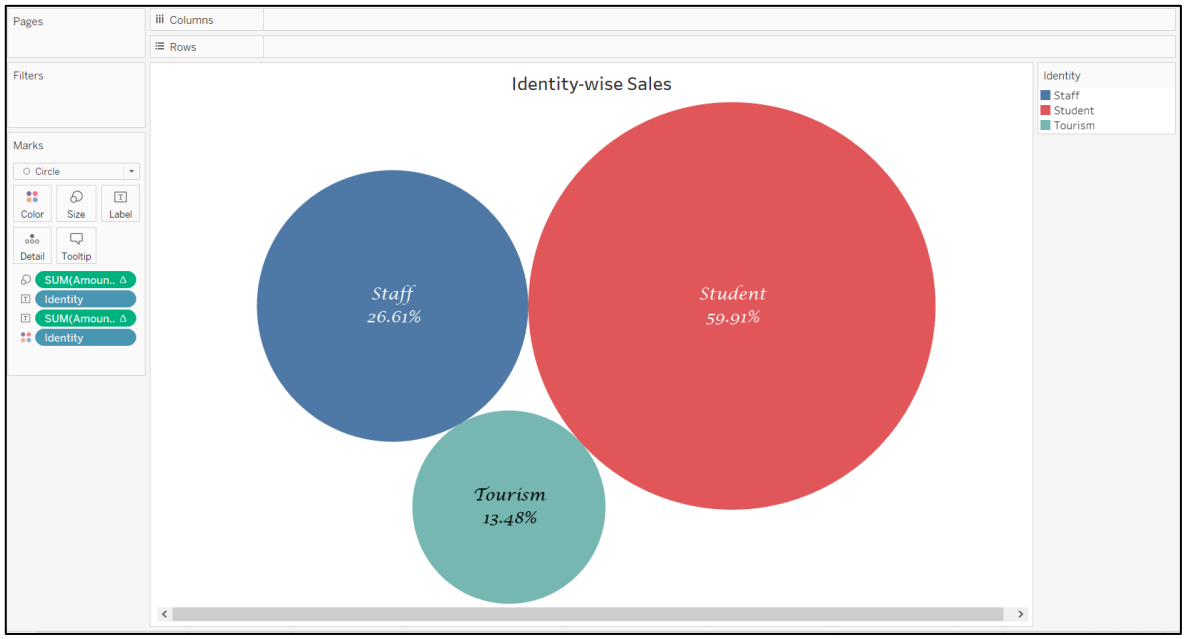
- R-Squared: 0.0071332
- P-value: 0.634857

Based on these values, we can conclude that our prediction is fairly accurate. Given below is the screenshot:



Ques. 2. Explore the sales performance of three consumer groups (students, staff and tourists).

Ans. Given below is the % of sales amount of identity-wise [Student has highest contribution with 59.91% of total sales amount]. The company's revenue, on the other hand, is reduced as a result of events. This should inspire a reconsideration of Happy Cow's events strategy, and someone should investigate whether or not efforts for that group should continue.



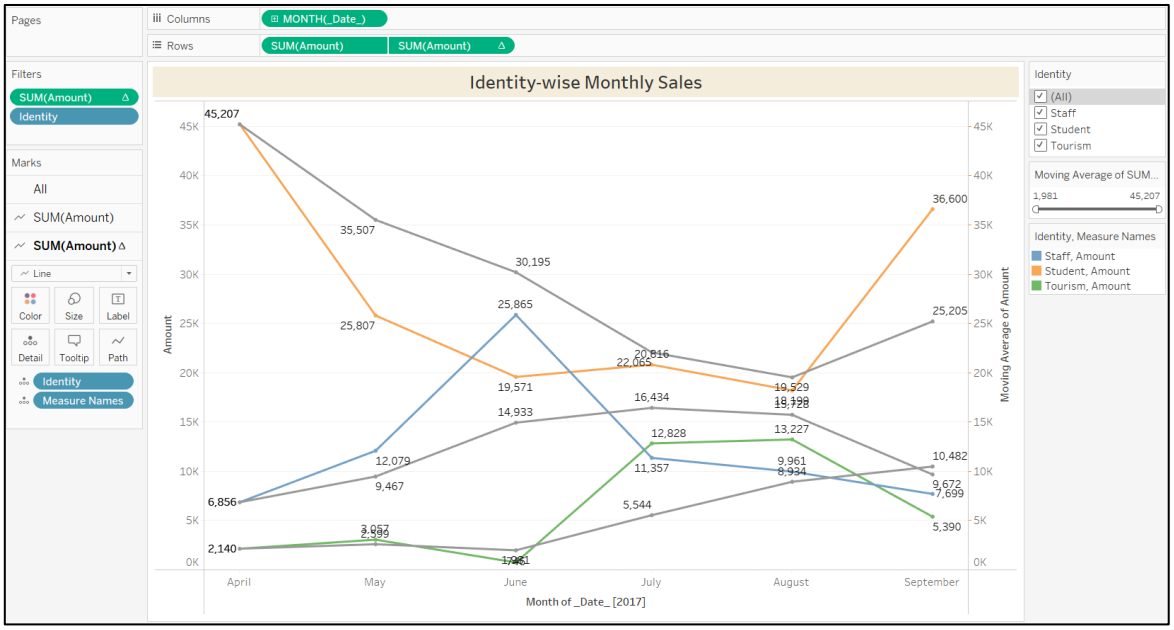
The following is a breakdown of monthly sales transactions by identity:

- The yellow line represents the student, while the grey line above it represents the moving average trend.
- The staff is represented by the blue line, and the moving average trend is represented by the grey line above it.
- Tourism is represented by the green line, while the grey line above it represents the moving average trend.

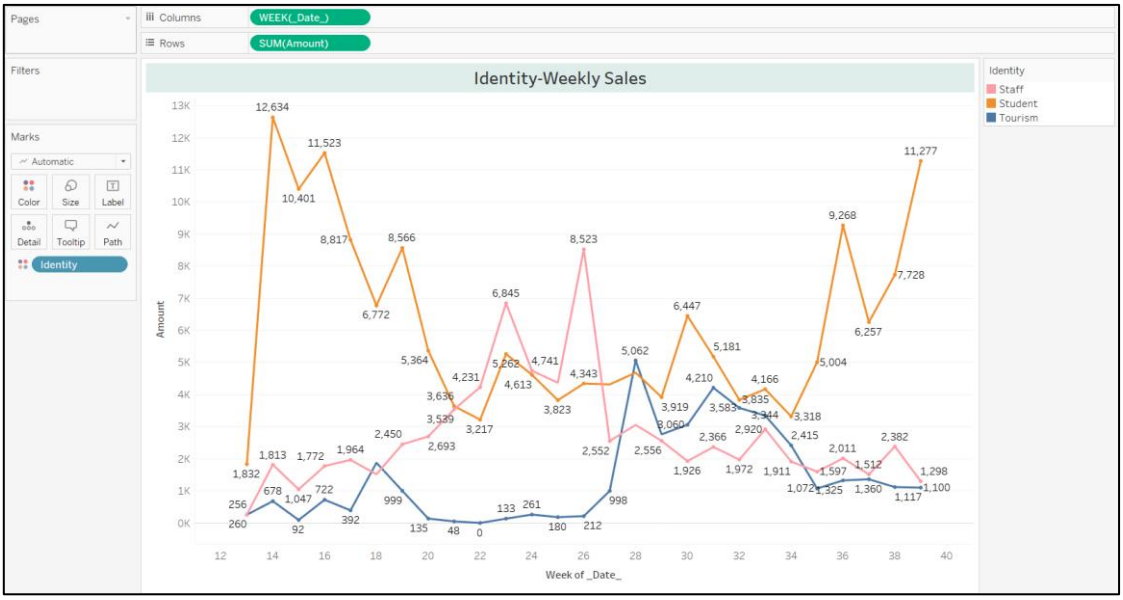
When we compare consumption groups month by month, we notice a shift in trends. The students deviate from the ice cream industry's regular trend of a definite preference during Hong Kong's hottest weather months, as indicated by a peak in April and early May, a drop in June, July, and August (rainy months), and then a breakout in September (month in which the rain ceased).

Even though student consumption was greater around week (the last week of June), staff consumption was higher, and tourist consumption was higher around week, according to the graph above (first week of July). It's possible that this is due to the summer break, since we've

noticed a gradual decrease in the weeks preceding up to it. Tourists were more common from the first week of July through the middle of July, presumably because people wanted to see how the campus looked before enrolling. It is visible from below graph:



Given below is the Weekly sales analysis of all the three Identities, Student's Sales trend is obviously high but were dropped right between week 20 to 29 and 32 to 35. Staff' Sales trend has a spike in between from week 22 to 27:



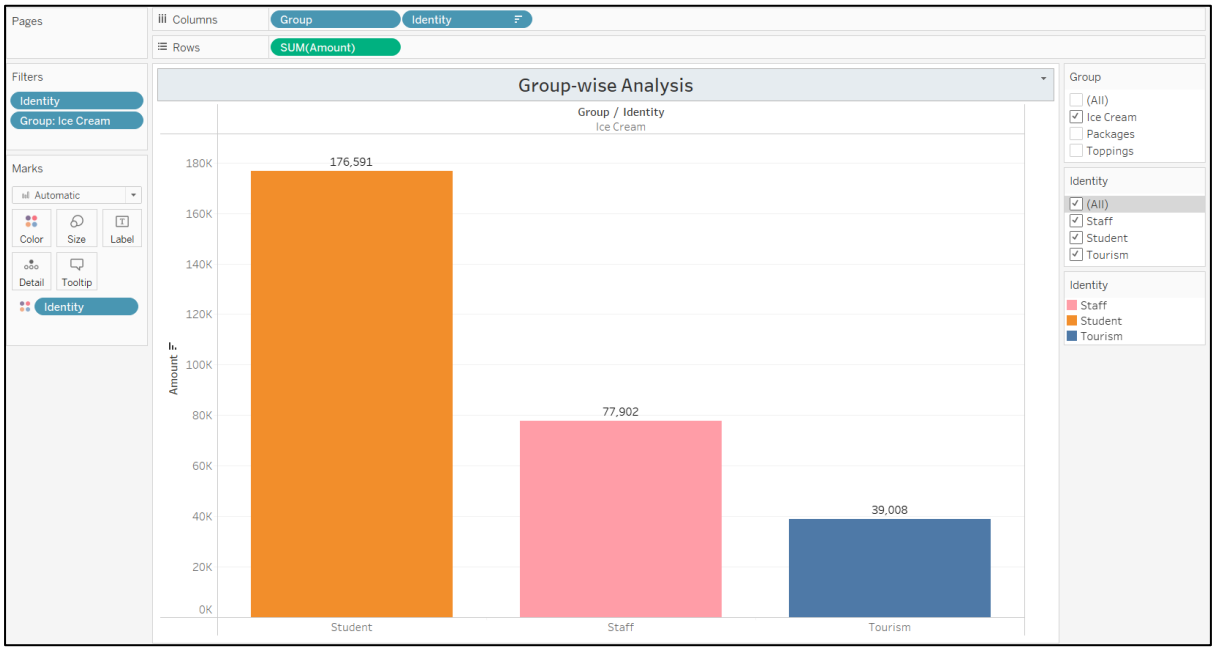
Ques. 3. Both Marry and Prem, senior sales assistants, believe that different groups of flavor sell better at different times of the year. Does the data back it up? Please propose your groupings and visualize them to generate insight into the ice cream sales. Regarding flavor groups, does grouping give a better level of analyses than individual level.

Ans. Yes, different groups of flavours sell better at different times of year and data is also there to support this. Before exploring the sales performance of three consumer groups there are few important identifiers which is required to explain, these identifiers has been used in further graphical representation:

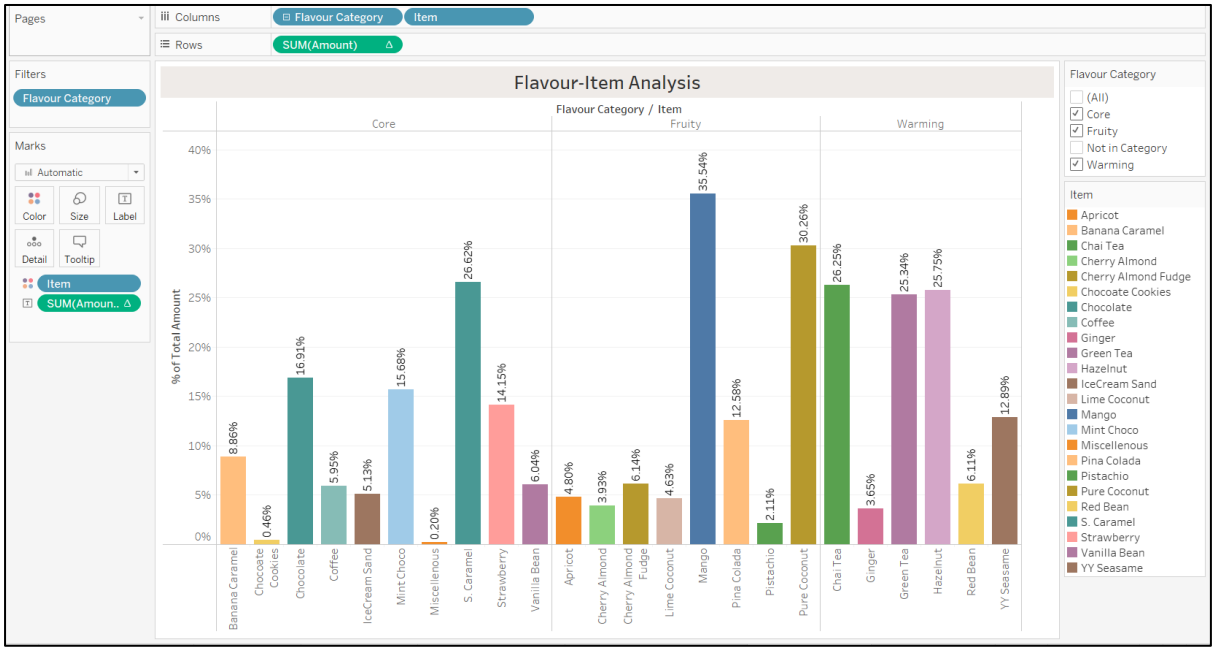
- **Identity** means the consumer groups i.e. Student, Staff and Tourists.
- **Flavour Category**: Fruity, Core, and Warming [Used to categories the Ice Creams]
- **Group**: Ice Cream, Package, and Toppings

Item	Flavour Category	Group
Banana Caramel	Core	Ice Cream
Chocoate Cookies	Core	Ice Cream
Chocolate	Core	Ice Cream
Coffee	Core	Ice Cream
IceCream Sand	Core	Ice Cream
Mint Choco	Core	Ice Cream
Miscellenous	Core	Ice Cream
S. Caramel	Core	Ice Cream
Strawberry	Core	Ice Cream
Vanilla Bean	Core	Ice Cream
Apricot	Fruity	Ice Cream
Cherry Almond	Fruity	Ice Cream
Cherry Almond Fudge	Fruity	Ice Cream
Lime Coconut	Fruity	Ice Cream
Mango	Fruity	Ice Cream
Pina Colada	Fruity	Ice Cream
Pistachio	Fruity	Ice Cream
Pure Coconut	Fruity	Ice Cream
1 Topping	Not in Category	Toppings
2 Topping	Not in Category	Toppings
3 Topping	Not in Category	Toppings
Double Scoop	Not in Category	Packages
Finger B 100	Not in Category	Ice Cream
Gift coupon 3 cone	Not in Category	Ice Cream
Gift Voucher	Not in Category	Ice Cream
Granola	Not in Category	Ice Cream
Purchase Gift Voucher	Not in Category	Ice Cream
Single Scoop	Not in Category	Packages
Stamp Card REDEEM	Not in Category	Ice Cream
Triple Scoop	Not in Category	Packages
Tub	Not in Category	Packages
Tub Ice Cream	Not in Category	Packages
Up 3 Toppings	Not in Category	Toppings
Waffle Cone	Not in Category	Ice Cream
Chai Tea	Warming	Ice Cream
Ginger	Warming	Ice Cream
Green Tea	Warming	Ice Cream
Hazelnut	Warming	Ice Cream
Red Bean	Warming	Ice Cream
YY Seasame	Warming	Ice Cream

Given below is the distribution shown for the group “Ice Cream” for all the three identities. Here Student is again at top with 176,591 as a total sales amount that is very high as compared to other two consumer groups:



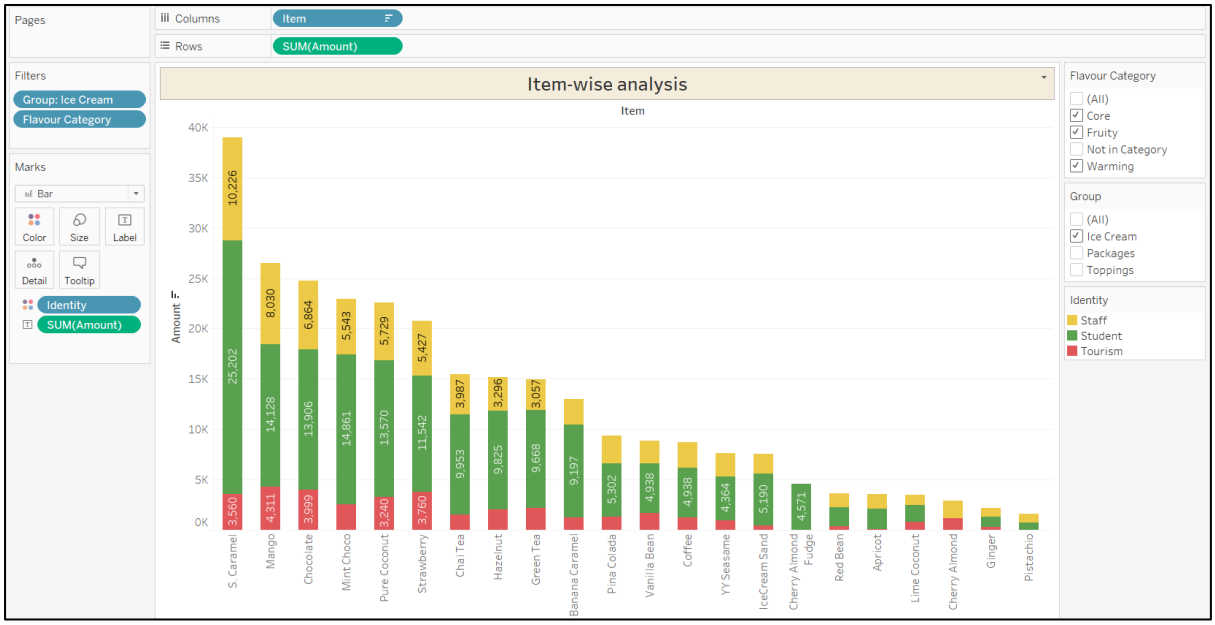
Given below is the Item-level analysis for each Flavour Category, here we can check the % contribution of each item for each Flavour. The Highest selling Items are again S. Caramel and Mango Core and Fruity, but Warming has 3-4 bestselling items.



Given below is the item-wise analysis for the three categories:

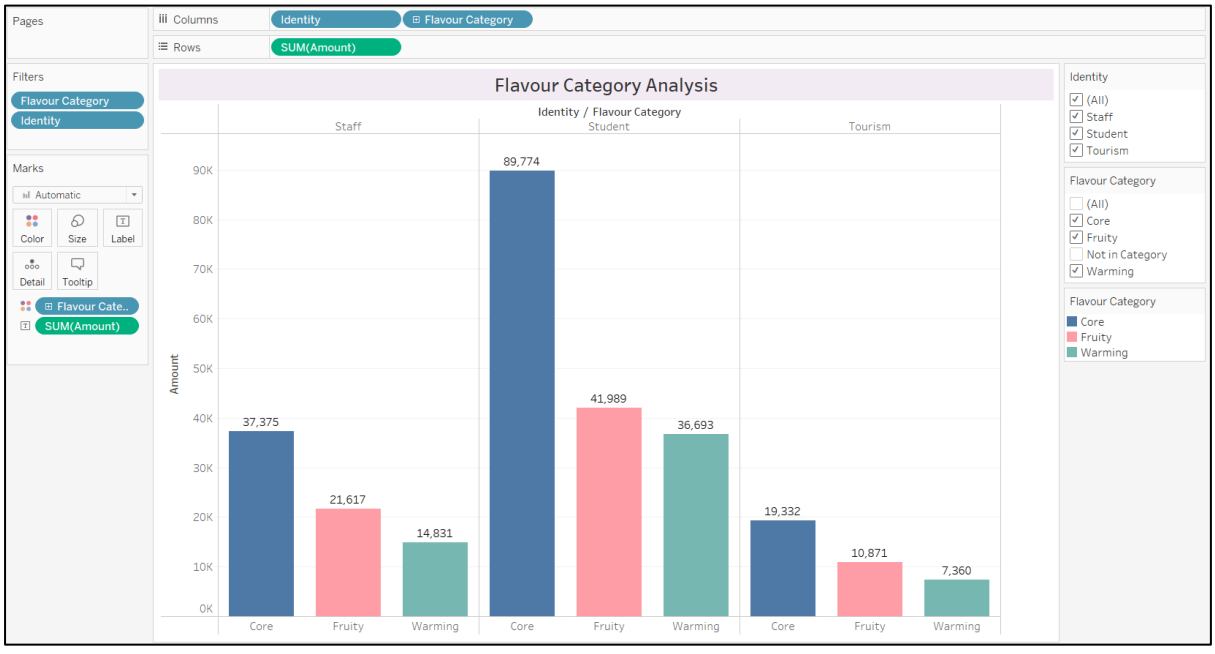
- Student's contribution is heavily visible in all the items.
- S. Caramel, Mango, Mint Choco, Pure Coconut, Strawberry are the top selling which are greater than 20K when accounted for 40k as the maximum sales amount.
- S. Caramel is sold the least in the month of August relative to other months but when you look at its contribution to the revenue of August month, we can observe that it still represents the second highest contributor in flavours (about 10.07% of the total revenue of August month). Thus, Happy Cow should maintain a good amount of stock for S. caramel flavour for all the months from April to September, that is including its lowest selling month of August.
- Mango is sold the best in the months of April, June and August relative to other months in the range of 5200 to 5500 units. However, despite its sale being much lower in the month of July compared to months of April, June and august, it still contributes the second largest (About 8.66%) to the revenue of the month of July. Hence, Happy Cow should maintain a good amount of stock for Mango flavour in July month in addition to April, June and August Months.
- Pistachio sells the best in the month of June (revenue as1468 units) compared to other months but when we look at its contribution to the revenue of June month, it accounts only about 2.88% of June month's revenue. Thus, Happy Cow should maintain low to moderate quantity of stock for Pistachio flavour in the month of June and should maintain very low stock of it for April, May, July, august and September months.
- Chocolate sells the least in the months of July (revenue as 2912 units) compared to other months in consideration. However, it is observed that its contribution to the revenue of July month is about 7.36% which is quite significant (4th largest contributor). Hence, Happy Cow should maintain high stock of chocolate flavour even in the month of July.
- Pure Coconut sales across months shows that it sells the second best in the month of April (revenue as 900 units). However, its contribution to the revenue of April month is very nominal, that is only 1.77% of the total revenue of April month). Therefore,

Happy Cow should maintain low stock of Pure coconut in all the months in consideration including April month and exception only being the June month (as it contributes about 4.59 %).

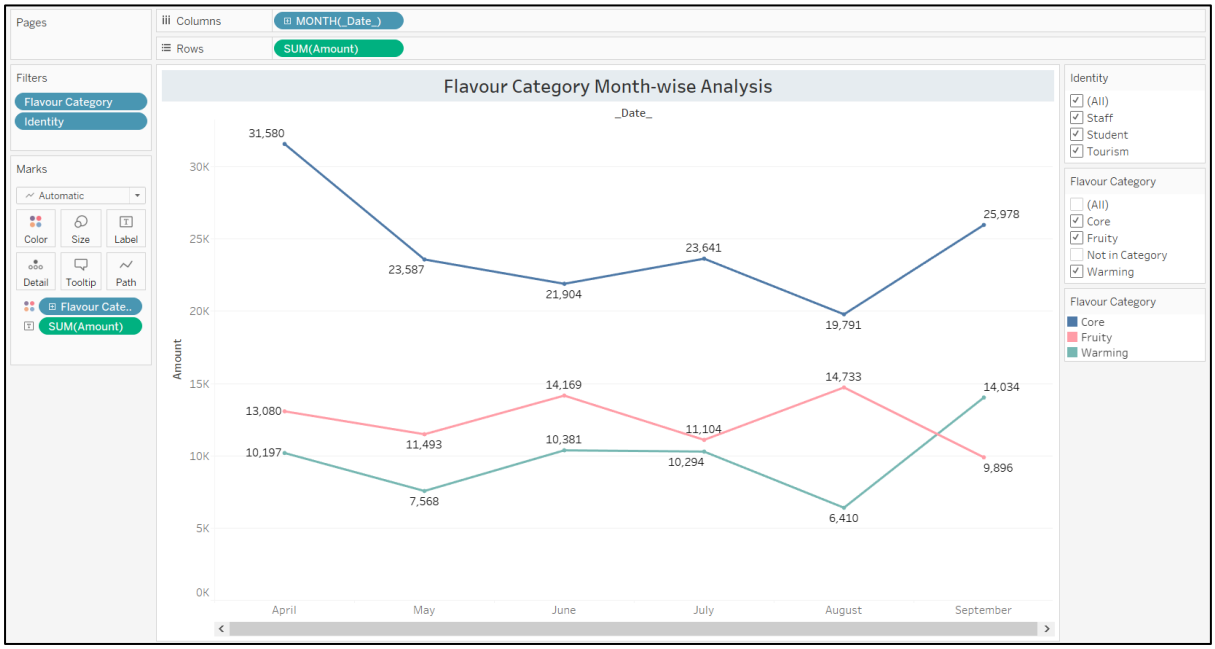


Given below is the analysis of Flavour Category vs Identity and their total sales amount, here we found:

- Student's sales amount is double or triple for each flavour.
- Core flavour is always on top for each identity.
- The least sold flavour is the "Warming".



Given below is the month-wise analysis of the flavour categories for all three identities [Student, Staff, and Tourism]. The "Core flavour is extremely behaving good and trend is high above the the other two. Warming Flavour has crossed the Fruity flavour but that happened in the month of September, which means the Seasonal Change:



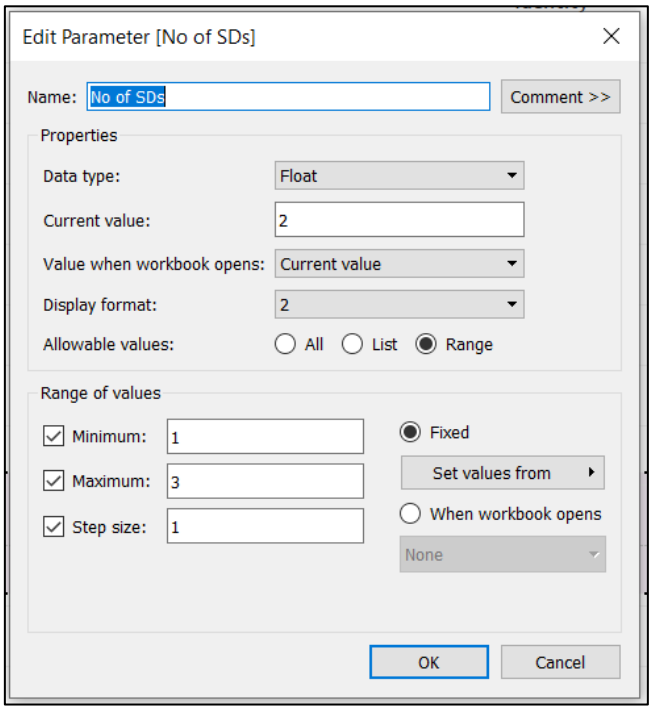
Ques. 4. What outliers can be identified from the daily sales of happy cow? Please define outliers and how to address them.

Ans. Outliers are values that are outside of the expected/accepted range, or that do not follow the trend and have a large divergence.

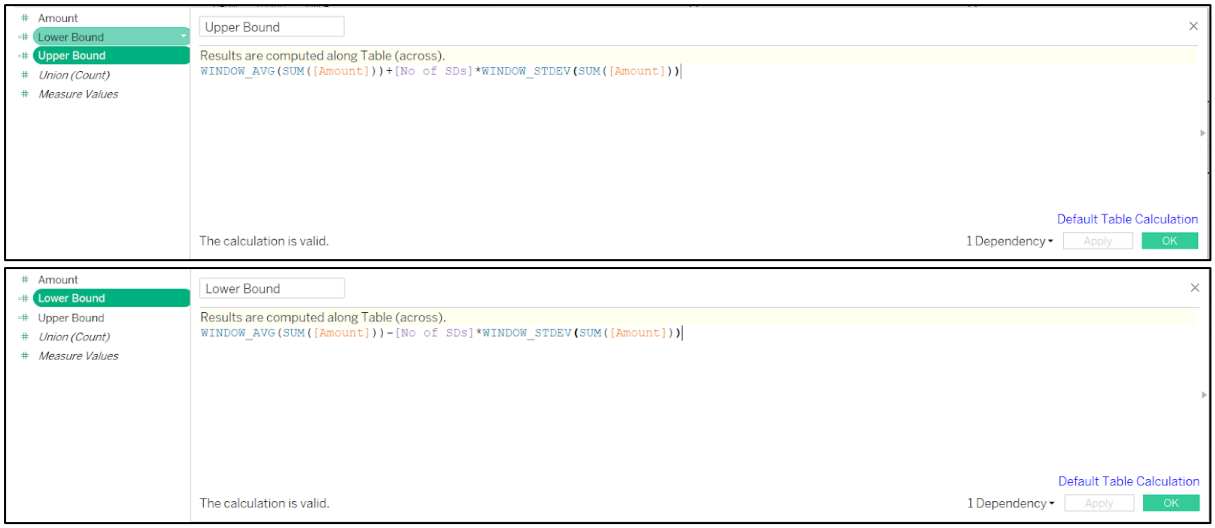
Student: When we try to visualize data of revenue generated by students over time and look at the graph, we can see that there are regular zero sales in the revenue, and after further analysis. We can see that the sales drop is on Sunday and a few other working days because there are no students, faculty, or tourists on Sunday.

Tourists: Tourists are a unique instance in which we have more negatives and null values since they prefer to utilize more coupons for sales, causing the graph to slant downward.

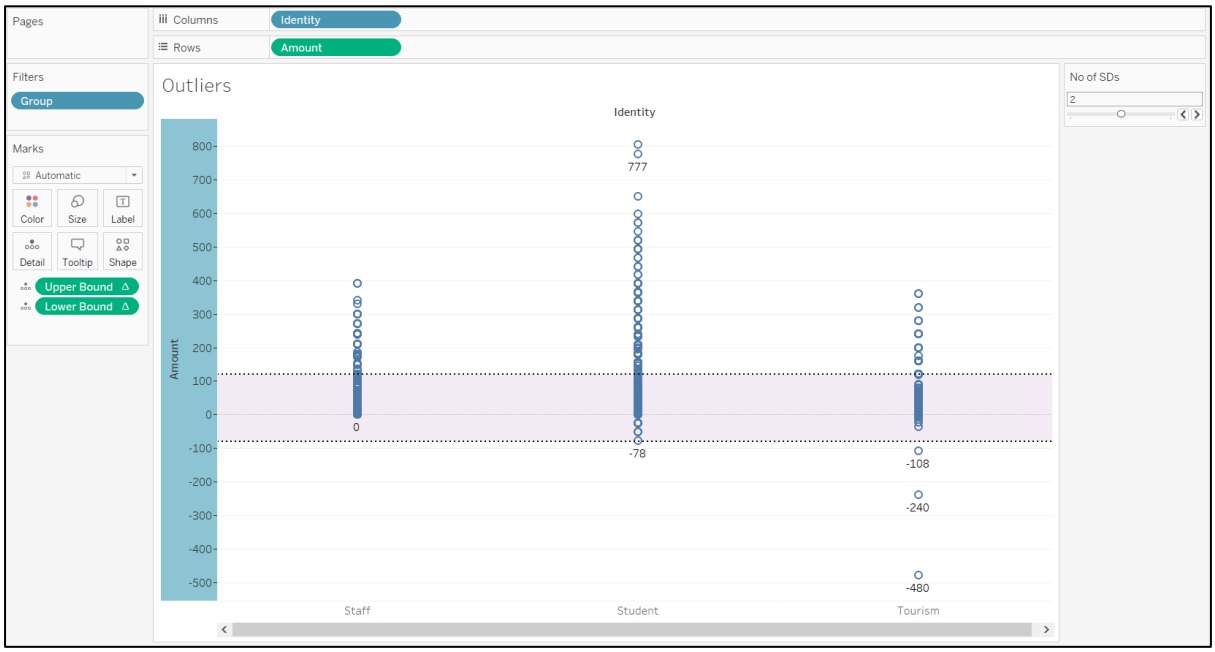
We first created parameter in Tableau to create “No. of Standard Deviation”, given below is the screenshot:



Then we created two calculated fields (a) Upper Bound and (b) Lower Bound, given below is the screenshot of it:



Given below is the screenshot of outliers taking 2 Standard Deviation:

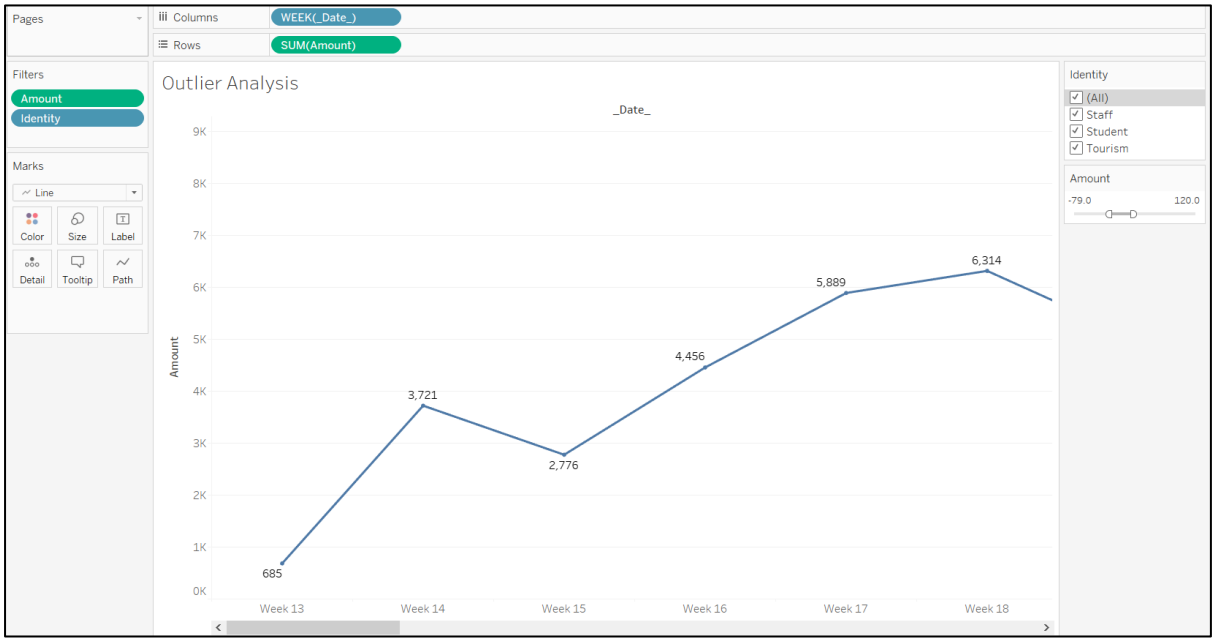


To work with outliers, I have identified below or all the three identities:

- Upper Bound: 119.4
- Lower Bound: -79.91

This gives the purview to visualize the outliers, so for Staff and Tourists the maximum amount of Items are falling in the range of Q1 to Q4 but the amount for Student is majorly falling in outliers.

Now, when analyzed the sales data within the Quartile range for all the three consumer groups then the trend shows a upward move right from initial weeks:



Following are the approaches to deal with the Outliers:

- **Predicting Confidence Level Approach**- Building a predictive model using historical data to estimate and obtain a feel of the overall common trend, seasonal or cyclic pattern of the time series data is one approach of accomplishing anomaly detection with time series data. We can create a confidence interval, or a confidence band, for the predicted values by using the predictive model to forecast future values and based on the error rates (which can be calculated using MAPE — Mean Absolute Percentage Error). Any actual data point that falls outside of this confidence band is an anomaly.
- **Unsupervised techniques** based on clustering are highly beneficial for anomaly identification since they do not require any annotated data indicating that a particular data point is anomalous. As a result, clustering methods can be particularly useful for detecting time series anomalies. Now, determining the number of clusters, which is required by most clustering algorithms as an input, is a common mistake or bottleneck for clustering algorithms for anomaly detection. Although there are numerous methods for calculating the number of clusters, dynamically predicting the number of clusters for each series is not possible with time series data.

Ques. 5. Give characteristics of Happy cow dataset? What are the purposes of time series analysis (i.e predictive versus descriptive)?

Ans. Following are the parameters that determine the characteristic of a dataset:

- **Dimensionality**- The number of properties that the objects in a data set have is referred to as the dimensionality of the data set. We have roughly 32 ice cream flavours in the Happy Cow dataset, as well as IDs classified as student, staff, or visitor. In addition, the attribute off time can be set on a daily or weekly basis. As a result, we may classify this dataset as having a high dimensionality.
- **Sparsity**- In some data sets, such as those with asymmetric features, the majority of an object's properties have zero values. This type of data is referred to as sparse data, or the data set possesses Sparsity. At the Happy Cow dataset, we see that the majority of the revenue on sales value is 0 in an hourly time frame. As a result, we can argue that this dataset is sparse.

Purposes of Time Series Analysis:

- **Forecasting**- What will the future look like based on the past and present? (and its uncertainty).
- **Smoothing**- What can I conclude about the true condition of nature in the past given a comprehensive (noisy) dataset?
- **Time Scale Analysis**- Determine which time scales of variation dominate or account for the majority of temporal variance in an observed set of data.
- **Filtering**- How should I update my estimate of the true condition of nature based on prior and current observations?
- What is the relationship between two time series of two phenomena, according to regression modelling.

Predictive vs Descriptive:

Predictive	Descriptive
Uses Historical data	Uses historical data
Fills in gaps in available data.	Reconfigure data into easy to read format.
Creates data models.	Describes the state of the business operations.
Forecasts potential future outcomes.	Learns from the past.
Answers 'What might happen'.	Answers 'What's happening'.

Click here to download the Tableau File:

https://drive.google.com/drive/folders/19F0sqLNm0pfzoSDAeoRdxo_AppziM5_s?usp=sharing