IMT 562: Exploratory Data Analysis

Pooja Sadarangani

About Data

For this Assignment, I have used the "Adidas Sales Dataset", which has been acquired from the Kaggle data repository.
Dataset Link:
https://www.kaggle.com/datasets/heemalichaudhari/adidas-sales-dataset

The dataset entails information regarding Adidas sales made in the year 2020 and 2021 and includes retailer information, types of products sold, location of sales, units of products sold, total revenue, and profit made from the sales.

This dataset can be used to gain insights into a plethora of aspects and assist with devising effective strategies and decision-making.

About Data

9650 Rows

14 Columns

Data Types: Num, String, Date

Data Transformation

The head rows of the original dataset were as shown below:

Adidas Sales Database

Retailer	Retailer ID	Invoice Date	Region	State	City	Product	Price per Unit	Units Sold	Total Sales	Operating Profit	Operating Margin	Sales Method
Foot Locker	1185732	1/1/20	Northeast	New York	New York	Men's Street Footv	$50.00	1,200	$600,000	$300,000	50%	In-store
Foot Locker	1185732	1/2/20	Northeast	New York	New York	Men's Athletic Foo	$50.00	1,000	$500,000	$150,000	30%	In-store
Foot Locker	1185732	1/3/20	Northeast	New York	New York	Women's Street Fc	$40.00	1,000	$400,000	$140,000	35%	In-store

When this dataset was directly loaded in R Studio and Tableau, it caused issues as the column headers took the value of Row 1, which is null in this case. To avoid this, I changed the formatting of the dataset as follows:

Retailer	Retailer ID	Invoice Date	Region	State	City	Product	Price per Unit	Units Sold	Total Sales	Operating Profit	Operating Margin	Sales Method
Foot Locker	1185732	1/1/20	Northeast	New York	New York	Men's Street Footwear	$50.00	1,200	$600,000	$300,000	50%	In-store
Foot Locker	1185732	1/2/20	Northeast	New York	New York	Men's Athletic Footwear	$50.00	1,000	$500,000	$150,000	30%	In-store
Foot Locker	1185732	1/3/20	Northeast	New York	New York	Women's Street Footwear	$40.00	1,000	$400,000	$140,000	35%	In-store
Foot Locker	1185732	1/4/20	Northeast	New York	New York	Women's Athletic Footwear	$45.00	850	$382,500	$133,875	35%	In-store
Foot Locker	1185732	1/5/20	Northeast	New York	New York	Men's Apparel	$60.00	900	$540,000	$162,000	30%	In-store
Foot Locker	1185732	1/6/20	Northeast	New York	New York	Women's Apparel	$50.00	1,000	$500,000	$125,000	25%	In-store
Foot Locker	1185732	1/7/20	Northeast	New York	New York	Men's Street Footwear	$50.00	1,250	$625,000	$312,500	50%	In-store

About Data

(Post Data Transformation)

9648 Rows

13 Columns

Data Types: Num, String, Date

Exploratory Data Analysis

If the dataset is not clean, the visualizations will essentially represent incorrect data and result in misleading insights. Thus, to begin with, I performed basic exploratory data analysis to ensure that I was working with a clean dataset.

My aim was to identify and remove any:

1. Missing values - There were no missing values in the dataset
2. Duplicates (rows) - There are no duplicate records in the dataset
3. Unreasonable values - There were no unreasonable values in the dataset

The above verifications were done manually and in R studio.

````
```{r}
# Checking for Missing Values
any(is.na(sales))
```

[1] FALSE

```{r}
# Checking for Unreasonable Values in each column
any(duplicated(sales))
```

[1] FALSE
````

Key Questions for EDA

01

Which Retailers made the highest and the lowest profits each year?

Which product contributed the most to the profits?
Did Amazon affect Walmart's Profits?

02

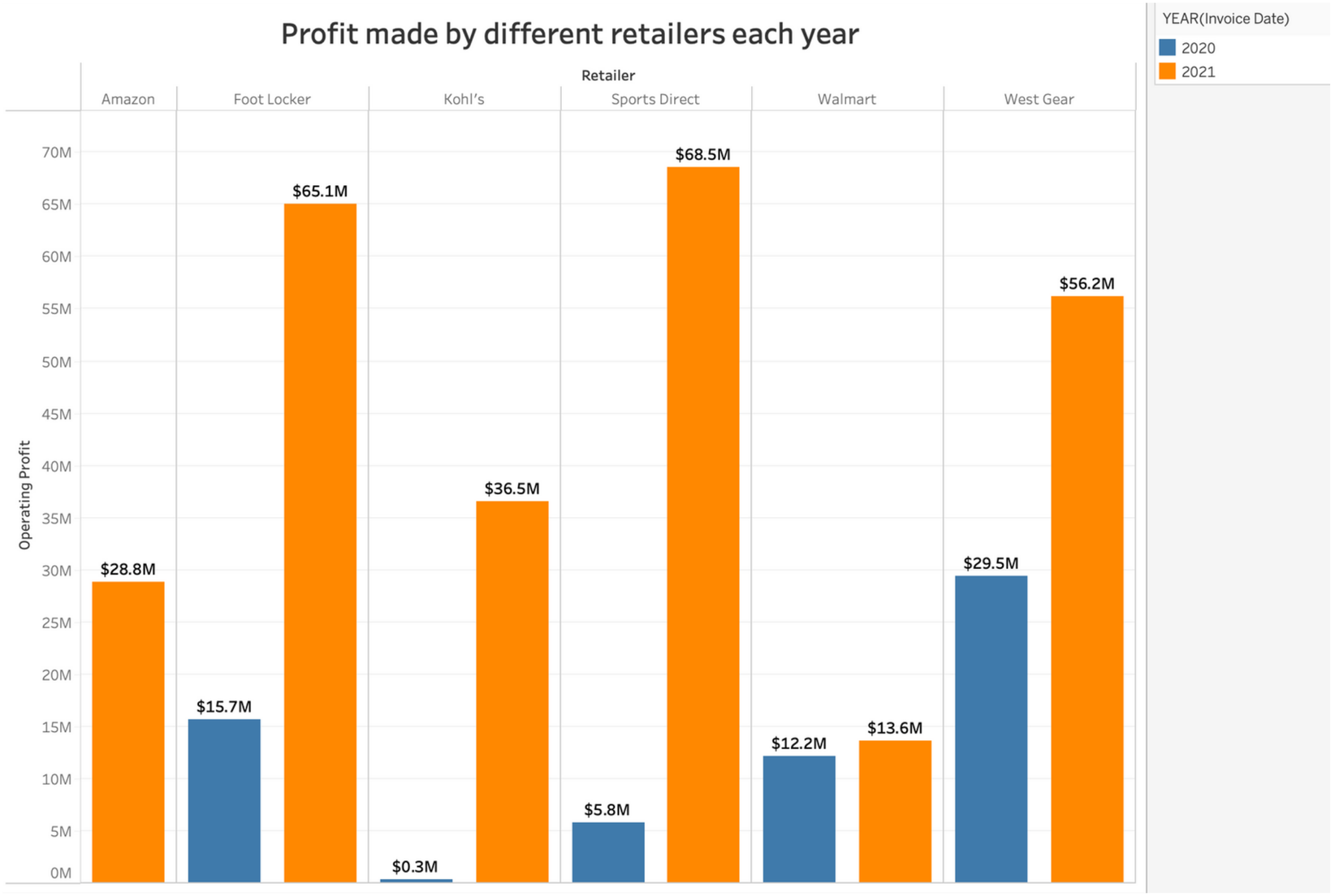
Which locations sold most number of units?

03

Which sales method is most popular amongst customers?

How does the popularity of a particular sales method vary with time?

Which Retailers made the highest and the lowest profits each year?

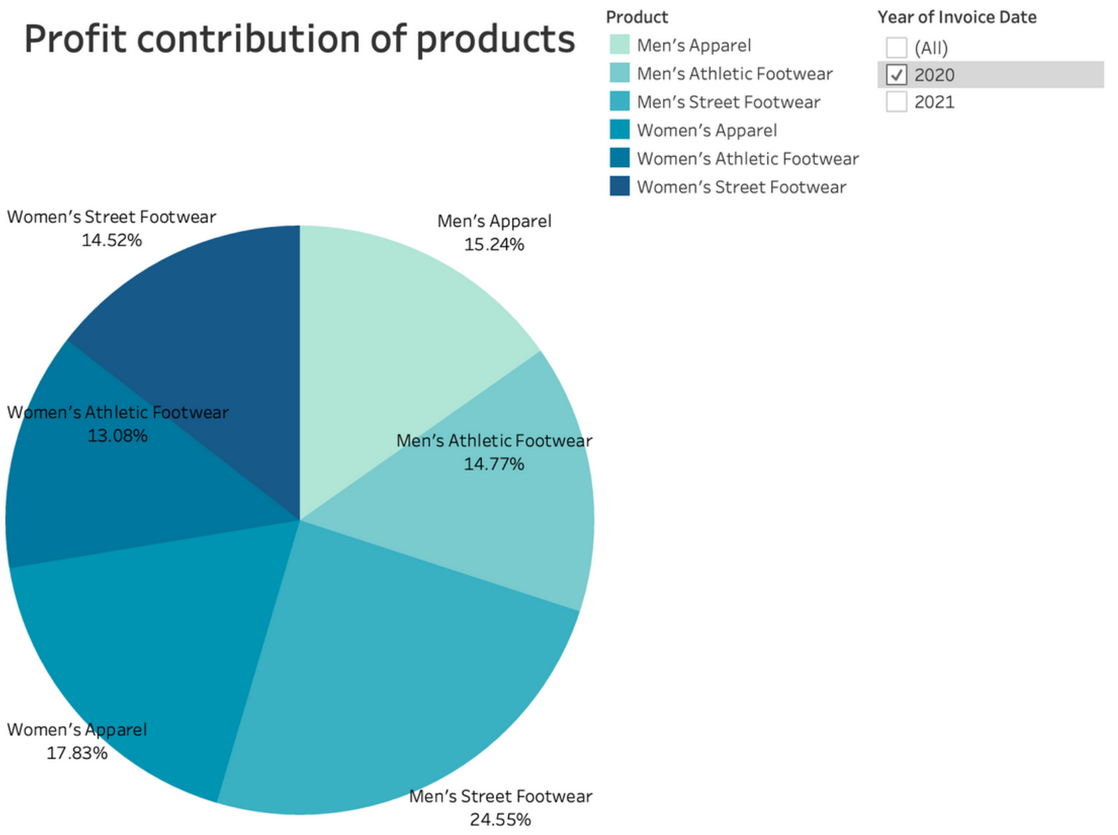


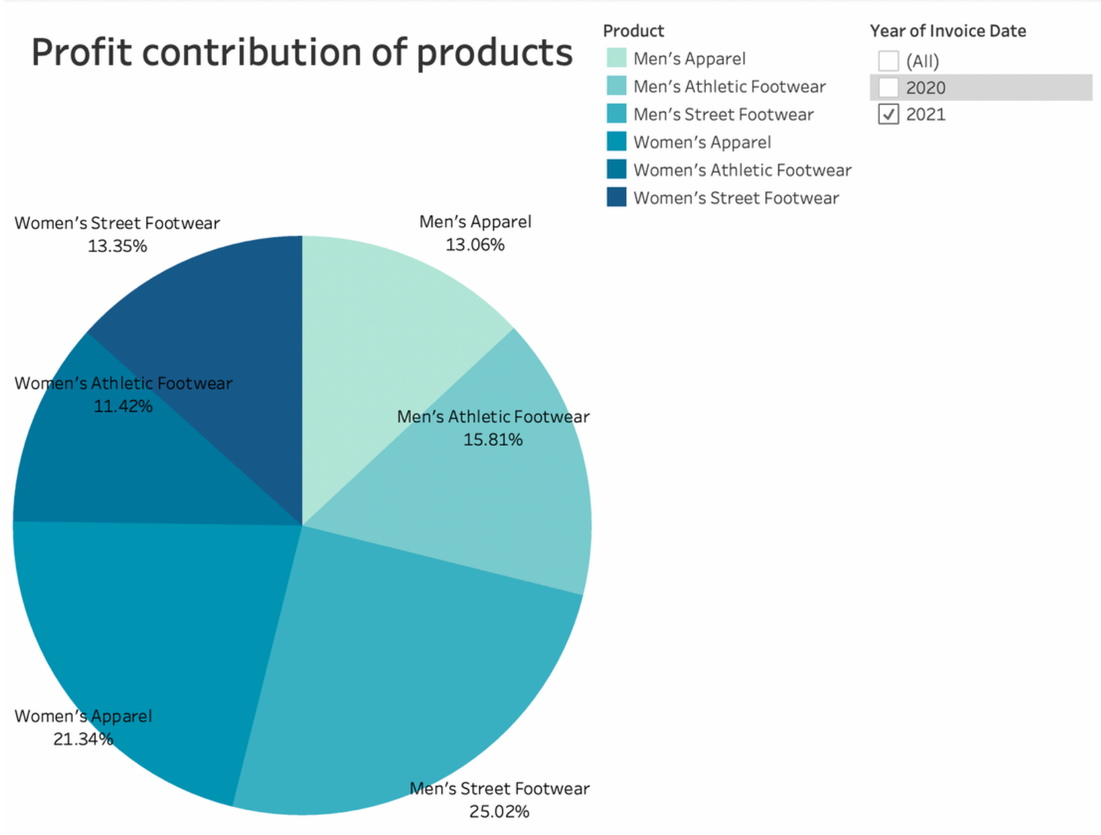
Key Insights

To answer the first key question, I have plotted a bar chart comparing the profits made by each retailer each year. Since there is a high difference in profit, instead of labeling the entire number, I have used 'Millions' as units to improve the readability.

1. In the year 2020, West Gear made the highest profits (approximately $29.5) and Kohl's made the lowest profits (approximately $0.3M). I am excluding Amazon from being considered for the year 2020 as there is no data for Amazon. This implies that Amazon did not sell Adidas products in 2020.
2. In the year 2021, Sports Direct made the highest profits (approximately $68.5) and Walmart made the lowest profits (approximately 13.6)
3. From this, we can also see that the profits have marginally increased for all retailers in 2021 except Walmart. We can further explore why this increment did not happen for Walmart. Did the entry of Amazon - Walmart's competitor - affect Walmart's profit?
4. Another question that arises is which products contributed the maximum to the profits in each year?

Sub-Question: Which Product contributed the maximum to the profits?





As per the pie chart, maximum profit was achieved from Men's street footwear in both years. We can understand that this product is consistently in high demand.

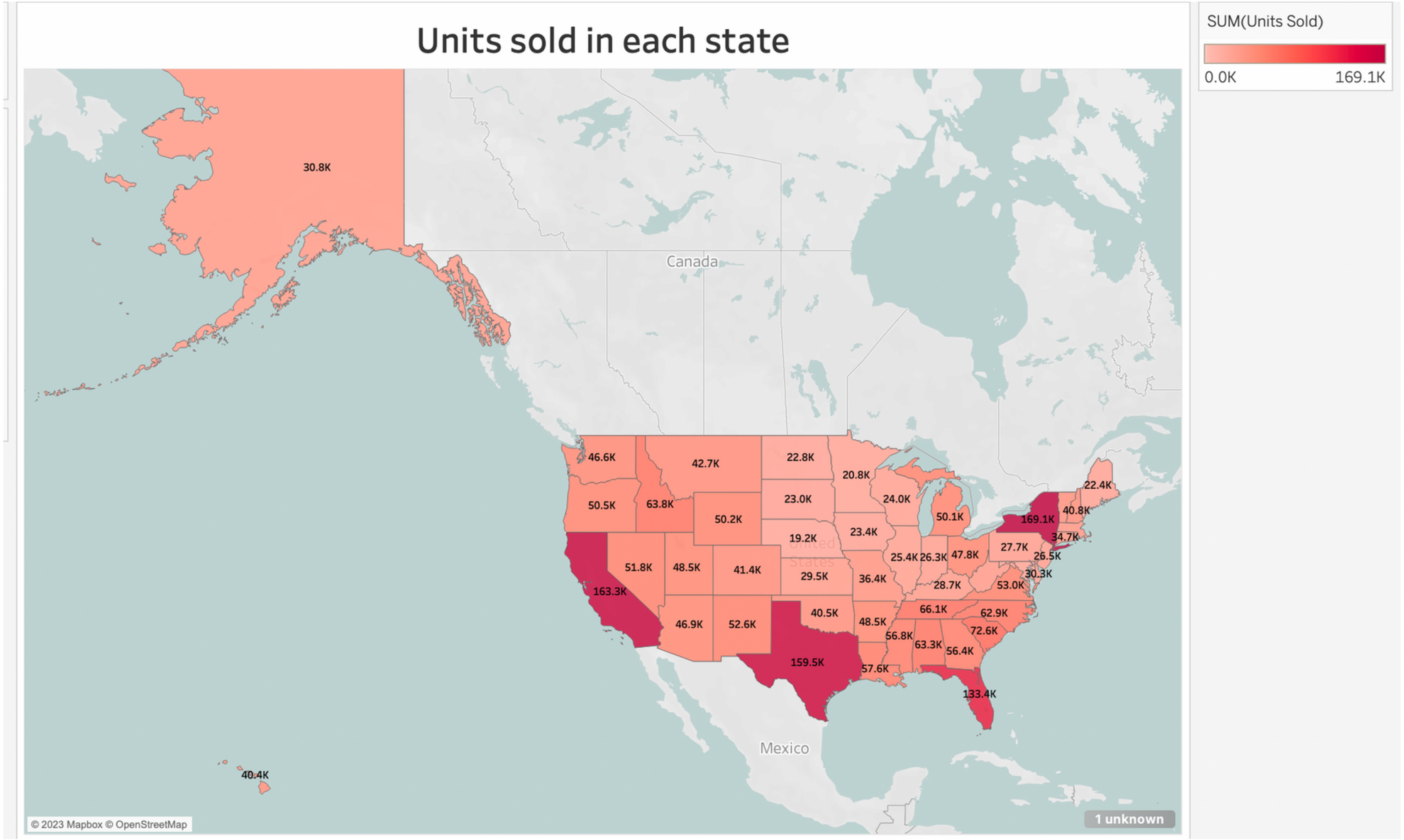
Sub-Question: Did the entry of Amazon - Walmart's competitor - affect Walmart's profit?

Data Transformation - Filtered data by Year=2021

This chart portrays the fact that Walmart profits are independent of Amazon. While it was logical to initially think that Amazon's entrance might have affected Walmart's potential profits, considering they are competitors, this chart tells otherwise.

We can clearly see from the chart that Amazon and Walmart profits have increased and decreased at the same time. This means that Walmart is not doing as well as others because of reasons other than the growing competition. .

Which locations sold most number of product units?



Which locations sold most number of product units?

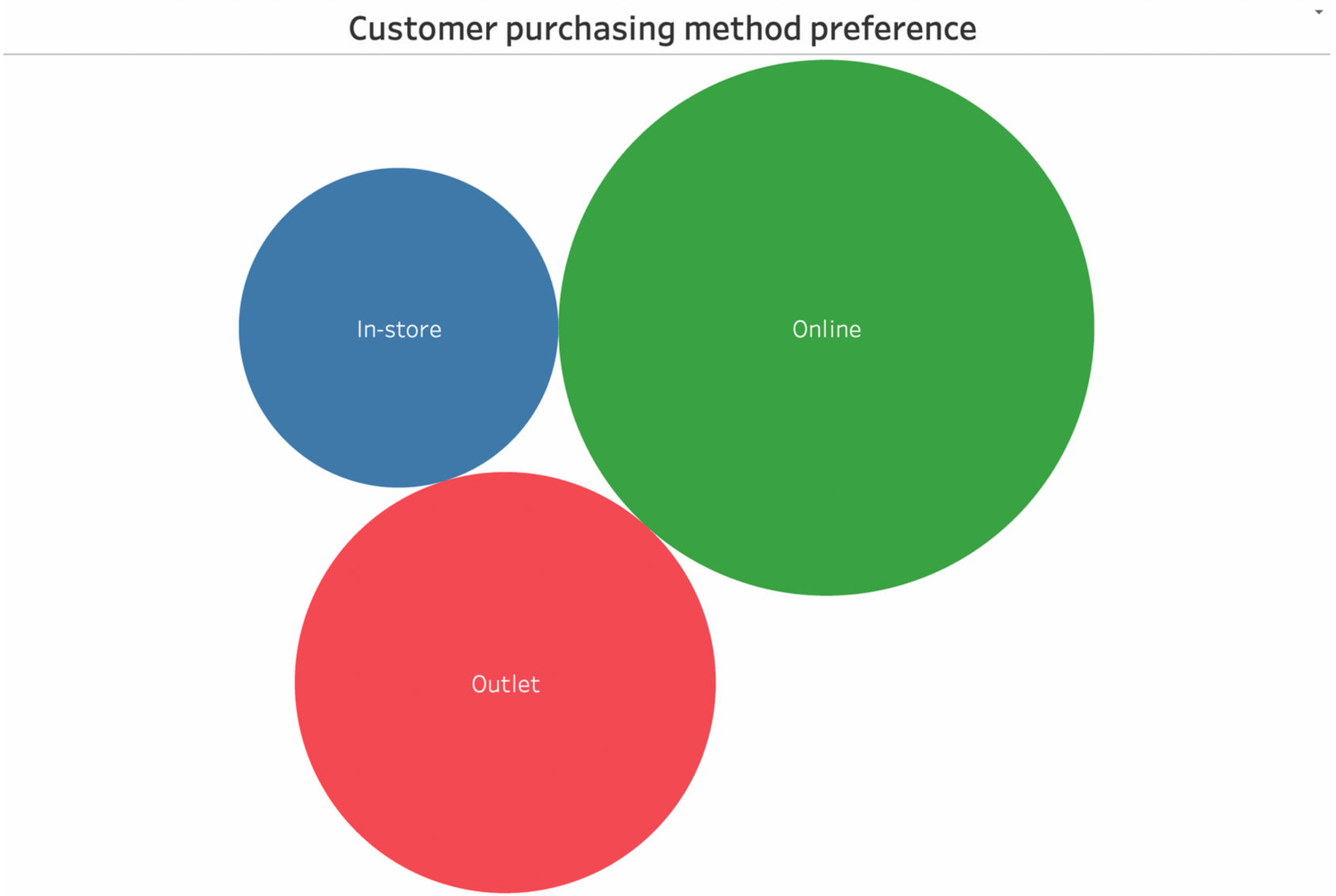
Key Insights

To answer the second key question, I have plotted a Map chart highlighting the states and the number of product units that were sold in each state. I added the units sold in the color mark to get the map. I used 'Thousands' units to label the quantity for improving the readability of the graph. I have also made another map chart to highlight the states that sold the most units of products. To do so, I clustered the data into 3 groups based on the units sold. Post this, I added the cluster in the filter and filtered the cluster of high units sold. I have labeled the states as well as the quantity. I have explicitly labeled the state despite this being a map because not everyone is well familiar with the US map.

The key takeaways from this chart are as follows:

A large number of units were sold in California, Texas, New York, and Florida. This makes sense because all 4 states are among the most populated states in the US. It is only logical that more products will be sold in locations that have a greater population.

Which sale method is most popular among customers.

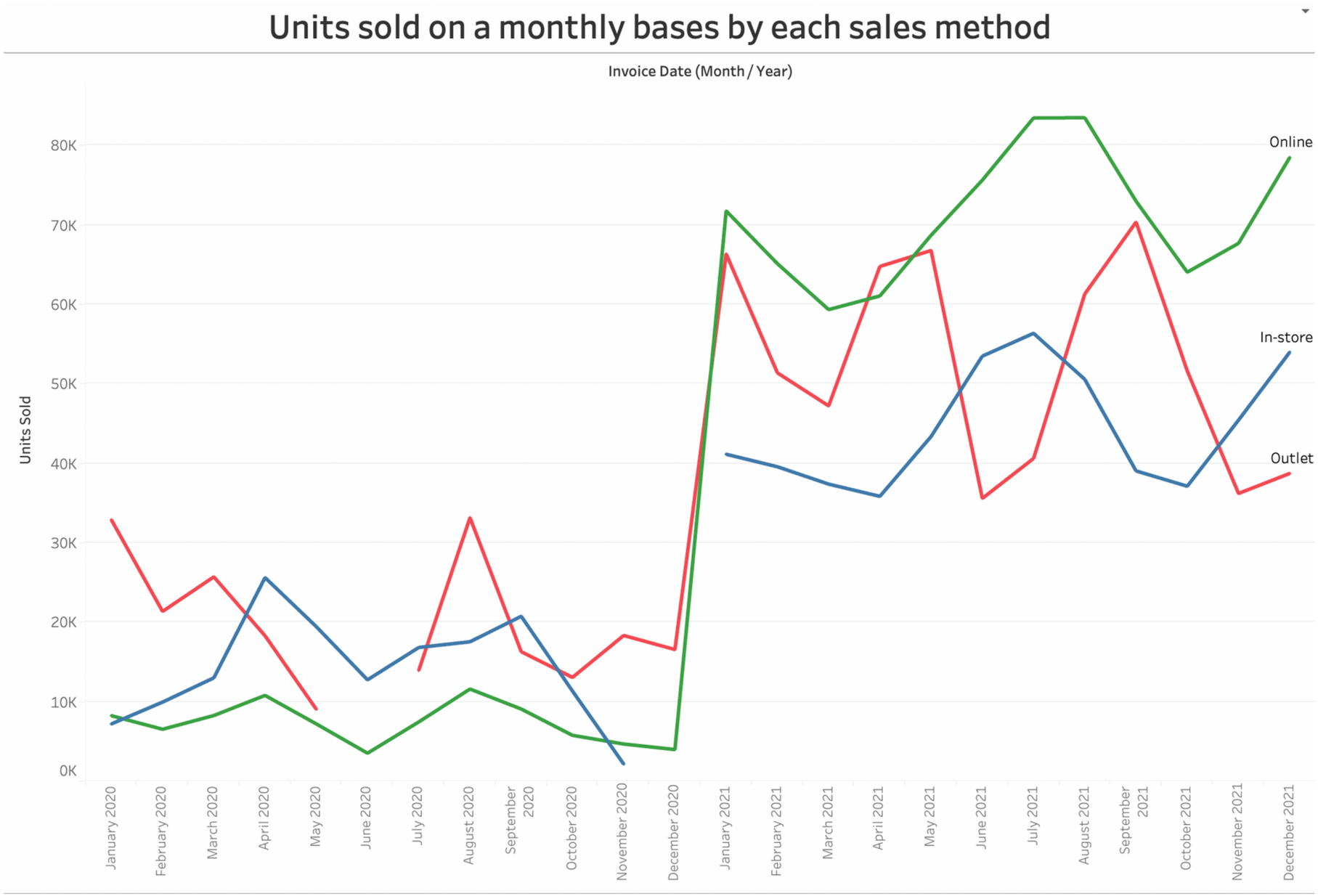


Key Insights

To answer the third key question, I have chosen a bubble chart. I chose the bubble chart solely because there was no need to know the metrics. The main aim was to determine which method is most popular. Some important insights from this chart are as follows:

1. The sales method preference is in the order: Online > Outlet > In-person. One of the core reasons for this order could be that online and outlet stores often have better discounts. Additionally, post covid, people have become accustomed to the online shopping experience and the benefits that come along with it such as less time investment.
2. Talking about offers brings us to the question of how the popularity of a sales method varies over a year. I have explored this sub-question in the next slide.

How does the popularity of a particular sales method vary with time?



Key Insights

For this, I have chosen a line chart that displays the number of units sold via each sales method on a monthly basis. The main aim was to determine any trends in the change throughout the year. Some important insights from this chart are as follows:

1. In the year 2020, not many units of Adidas products were sold. However, this changed in 2021. This could be because of the pandemic. In 2020, since many people lost their jobs, they were spending money on only necessities. However, in 2021 when things started to go back to normal, people were more willing to buy other products.
2. The quantity sold peaks end of every season, possibly due to discounted prices as part of "End of the season sale"
3. Another aspect to note is that online shopping was the least preferred method in the year 2020, but the most preferred method in the year 2021. Again, this could be due to the pandemic. People were exposed to the online shopping world and got accustomed to online shopping and the benefits that come along with it such as convenience and less time consumption.

Conclusion

01

Sports Direct made highest profit in year 2021 and Walmart made the lowest profit in year 2021.

02

Unlike other retailers, Walmart profits did not hike in the year 2021.

03

Customers now prefer online shopping over other modes of shopping

04

Most units of Adidas were sold in California, Texas, Florida, and New York locations.

05

The number of products sold increases highly at the end of every season.

Future Exploration

It would be interesting to integrate this dataset with customer data and gain more insights into customer purchasing behavior to understand why some retailers, locations, and sales method are doing better than others.