

**Exploratory Data Analysis**

On Kaspersky Antivirus Sales, August 2019

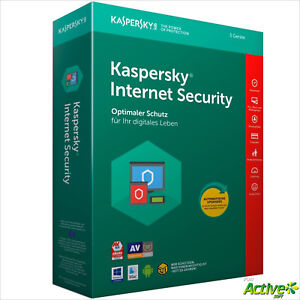
24.10.2022

**Submitted By**:

S. Mohammad Saif Rizvi – A22028

Tanushree Mandal – A22038

Pallavi Mazumdar – A22025

**DS August 2022**

**STS Project**

***“This highly confidential data presented in this report is only for research and analysis purpose and not be collected or shared by any means to avoid any kind of judicial activities.”***

**Motivation behind the Project**

The aim of this project is to understand the performance of Kaspersky in India and gain various insights on how this company is dominating the Indian market with its internet security products using Exploratory Data Analysis.

In 2019, Kaspersky was reportedly the #1 antivirus software according to Bisinfotech, followed by its competitors – Norton Security, Trend Micro, Quick Heal, Microsoft Security Essentials, ESET, Bitdefender, AVG, BullGuard and Panda Security. In this project we are specifically studying the Sales of Kaspersky antivirus products via Amazon for August 2019; State wise, City Wise and Postal Code Wise to understand which is the top contributor to the Sales of Kaspersky products.

**Introduction**

Kaspersky is a Russian multinational cybersecurity and anti-virus provider headquartered in Moscow, Russia, and operated by a holding company in the United Kingdom. It was founded in 1997 by Eugene Kaspersky, Natalya Kaspersky, and Alexey De-Monderik; Eugene Kaspersky is currently the CEO. Kaspersky Lab develops and sells antivirus, internet security, password management, endpoint security, and other cybersecurity products and services.

Kaspersky expanded abroad from 2005 to 2010 and grew to $704 million in annual revenues by 2020, up 8% from 2016, though annual revenues were down 8% in North America due to U.S. government security concerns. As of 2016, the software has about 400 million users and has the largest market-share of cybersecurity software vendors in Europe. Kaspersky Lab ranks fourth in the global ranking of antivirus vendors by revenue. It was the first Russian company to be included into the rating of the world's leading software companies, called the Software Top 100 (79th on the list, as of June 29, 2012). Kaspersky Lab is ranked 4th in Endpoint Security segment according to IDC data for 2010. According to Gartner, Kaspersky Lab is currently the third largest vendor of consumer IT security software worldwide and the fifth largest vendor of Enterprise Endpoint Protection. In 2012 Kaspersky Lab was named a "Leader" in the Gartner Magic Quadrant for Endpoint Protection Platforms.

According to WIRED, Kaspersky's software was "advanced for the time". The company's revenue grew 280 percent from 1998 to 2000, with about 60 percent of its revenue coming from foreign sales. It opened offices in the UK, Poland, Holland and China. It later expanded to Germany, France, the US and Japan. By 2000, the company had 65 employees and sales in more than 40 countries. Kaspersky opened new offices in South East Asia and the Middle East in 2008 and in South Africa in 2009. It also expanded in India, the Middle East and Africa in 2010. In 2009, retail sales of Kaspersky Lab's antivirus products reached almost 4.5 million copies per year.

**Methodology**

* Data Source & Description

This data is taken from SpeedTech Company and contains 7 columns and 12,267 records.

The columns included in our analysis are described below in brief:

1. Order Date – This field contains the date on which the customer placed the order on Amazon.
2. Order ID – This field contains the ID of the product that was ordered.
3. Quantity – This field contains the quantity of product that was ordered.
4. City – It lists the Names of the City from where the order was placed.
5. State – It tells us the Names of State from where the order was placed.
6. Postal Code – It contains the PIN code of the address from where the order was placed
7. Sales – It contains the Total Order Amount (in Rupee) of the product that was ordered.

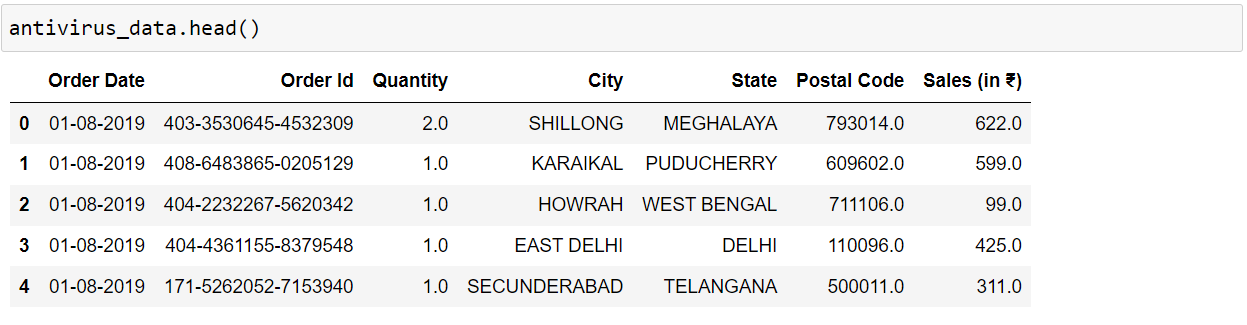
* Tools Used
* Python (Jupyter)
* Tableau
* Microsoft Word

**Analysis**

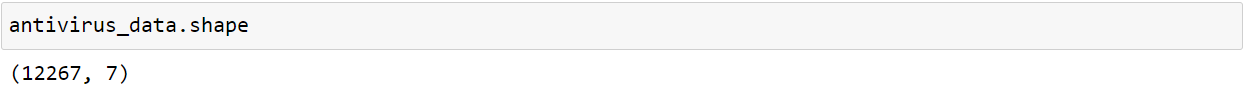
* Data Cleaning & Pre-processing

Importing the Pandas and Numpy libraries and loading the data

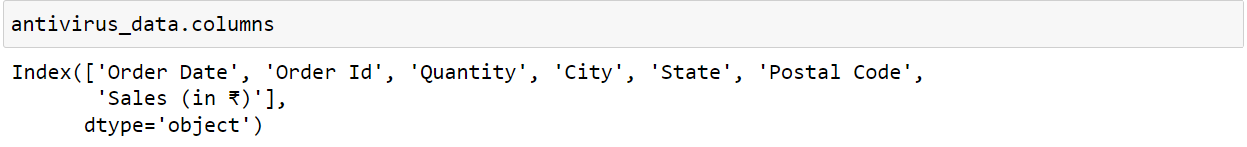




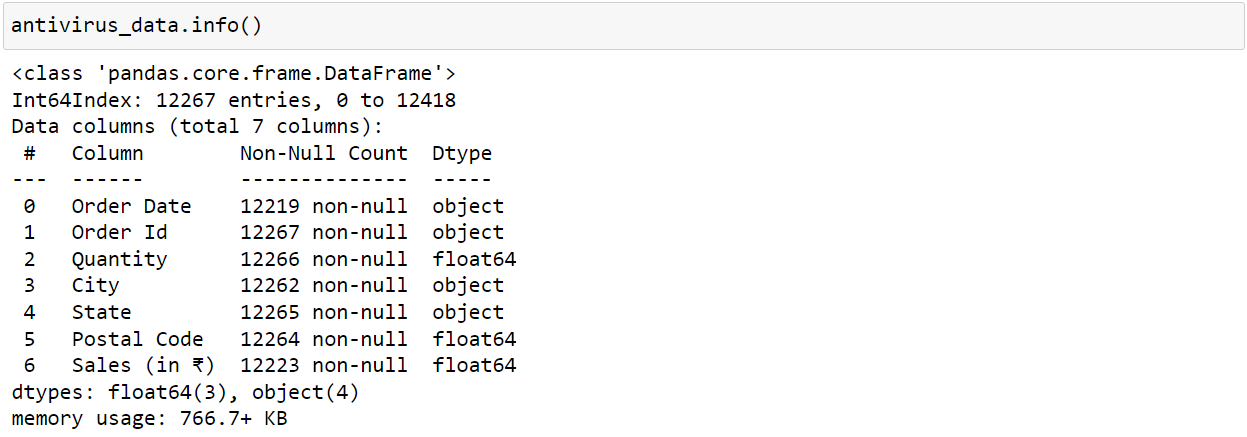
Displaying the Number of Rows and Columns of our data using ‘.shape’



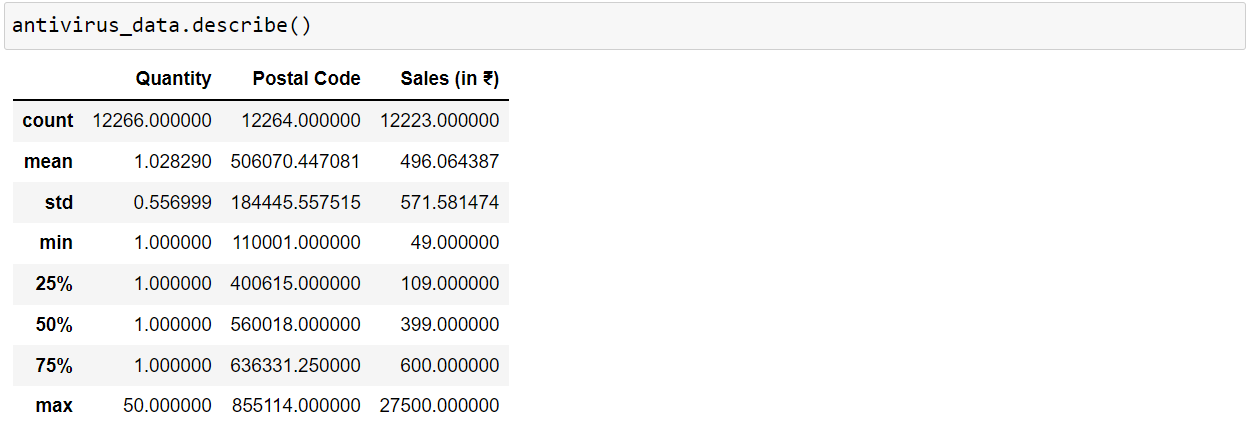
Displaying the Name of the Columns in our data using ‘.columns’



Checking the Data type and the number of Non-null Values in each column using ‘.info()’. Here we can see that except for ‘Order ID’, we have Null values for all the other column which we’ll have to deal with.

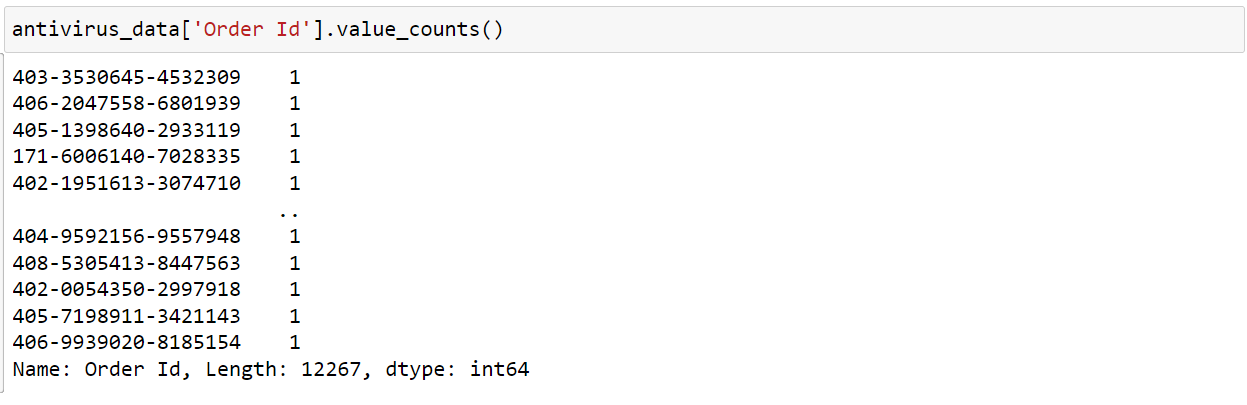


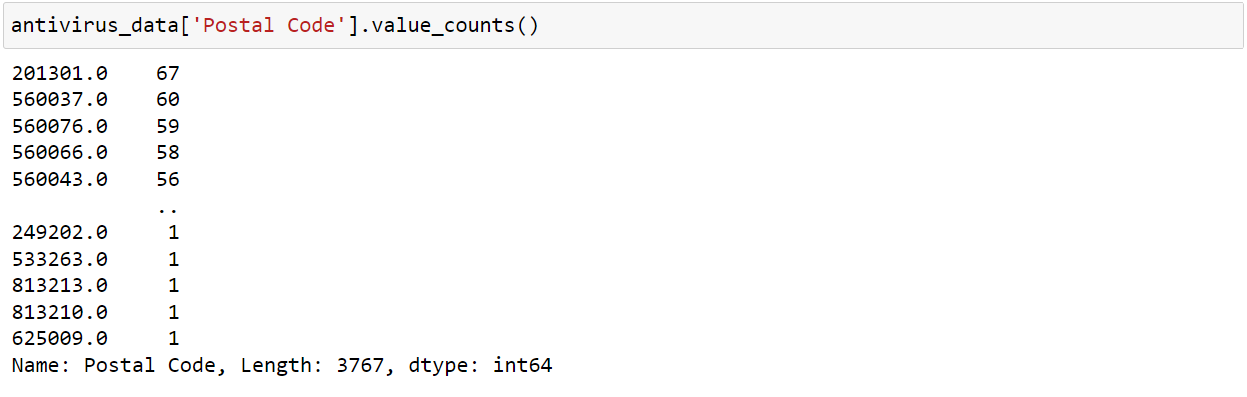
Using ‘.describe()’ to get the 5 point summary of the numerical columns. Since ‘Postal Code’ has an integer data type, Python displays the Descriptive statistics of this column too.



Using ‘.value\_counts()’ we can get the count of each column’s unique records. Here we can see that for the column ‘Order Date’, we have maximum number of records for the date 11-08-2019 signifying we have the most Sales for this date.



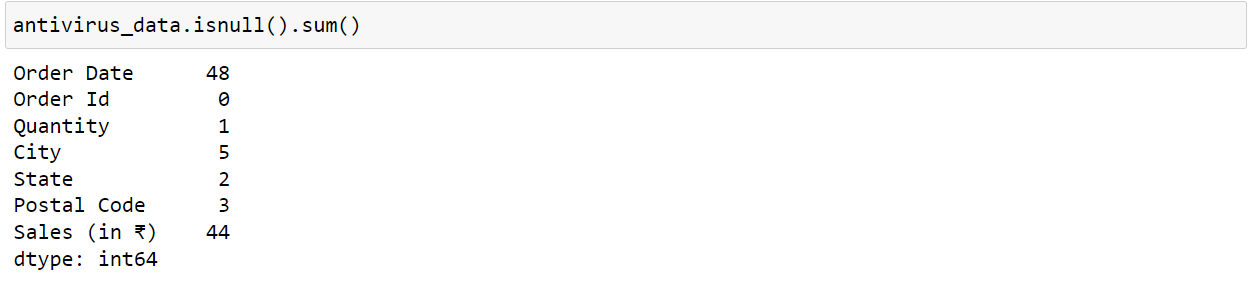




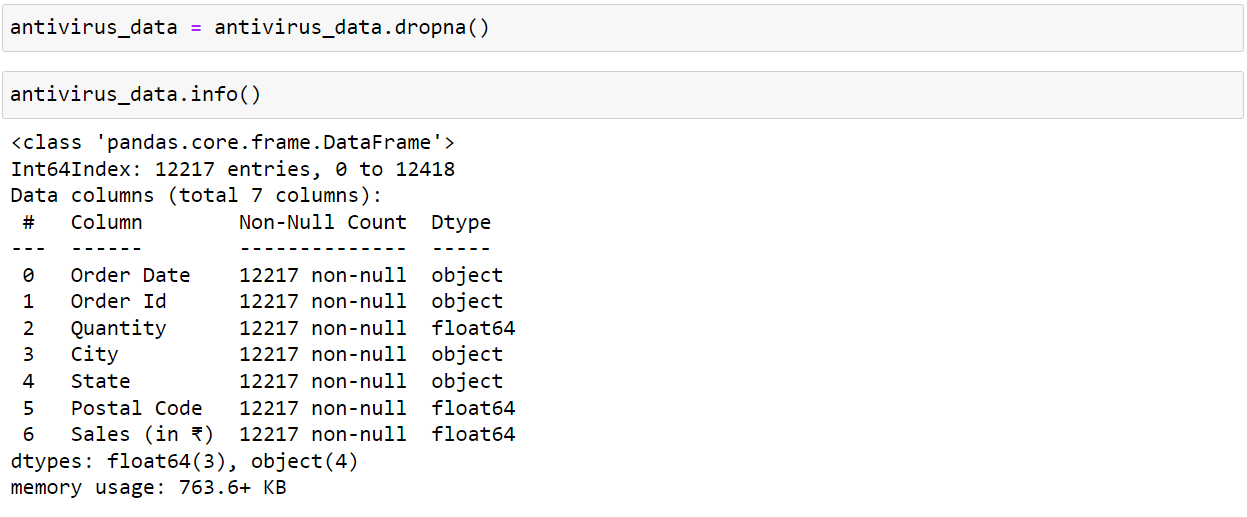
Here we are checking if our data contains any null records using ‘.isnull()’, but as we can see its only displaying the first 5 records and last 5 records

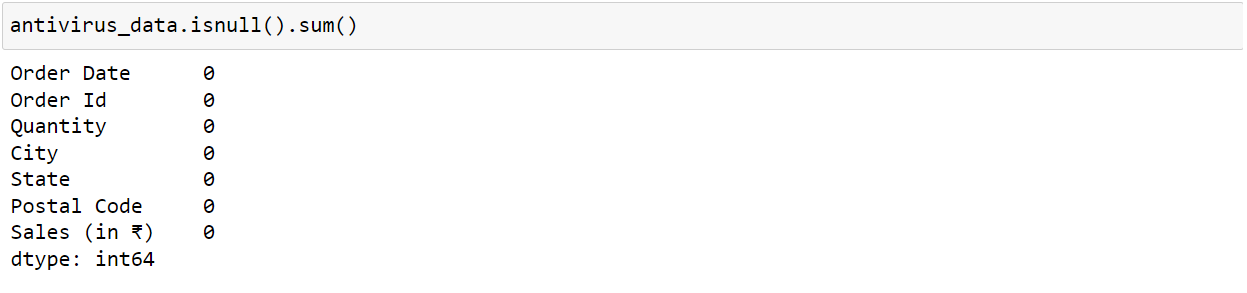


So instead of using just the ‘.isnull()’ method, we add ‘.sum’ to it in order to get the total number of null values for each column and as we can see, we have 48 null values for ‘Order Date’, 44 for ‘Sales’, 5 for ‘City’, 3 for ‘Postal Code’, 2 for ‘State’ and 1 for ‘Quantity’



Since there are very few missing values (<5%), we will drop all the missing values using ‘.dropna()’. After dropping the null values we are now left with 12,217 records.



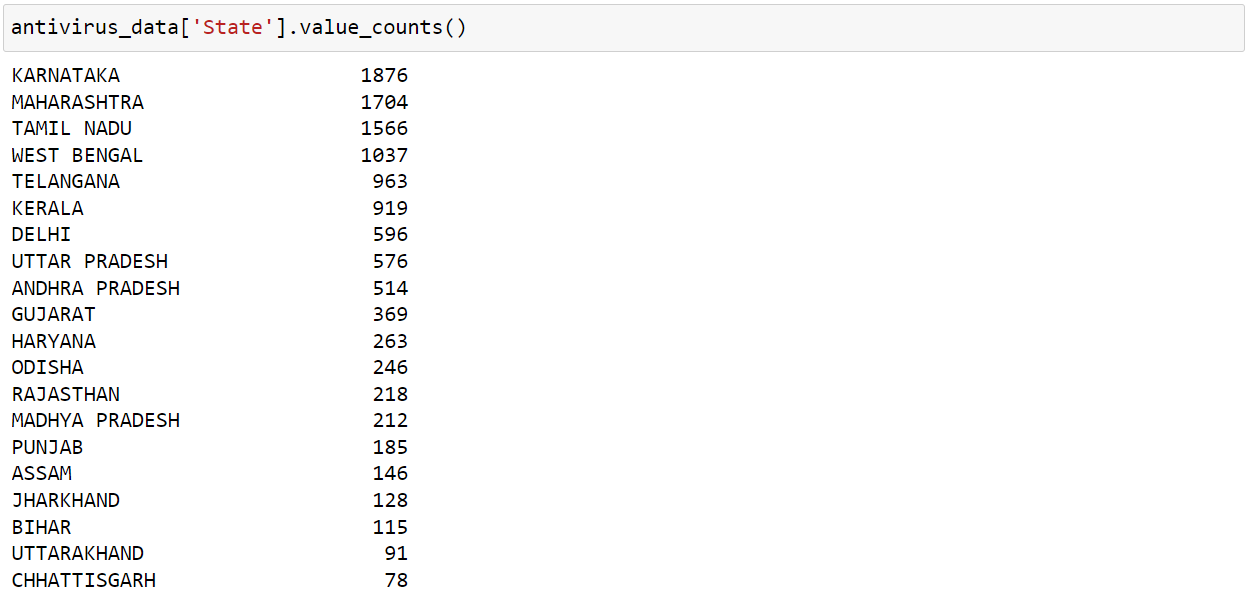


Since there were quite a few records under ‘State’ and ‘City’ columns where some of the names were in lowercase while some in uppercase, we are converting all of them into uppercase using ‘.str.upper()’

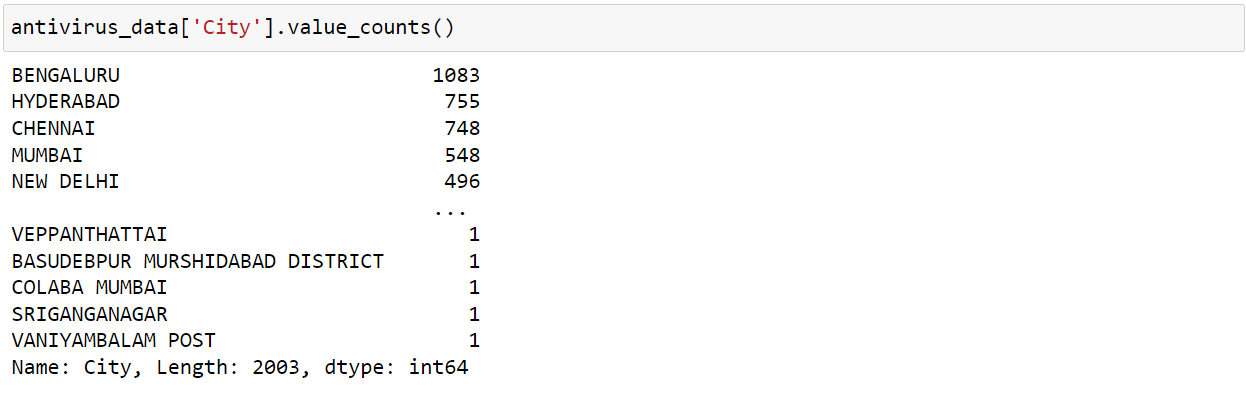


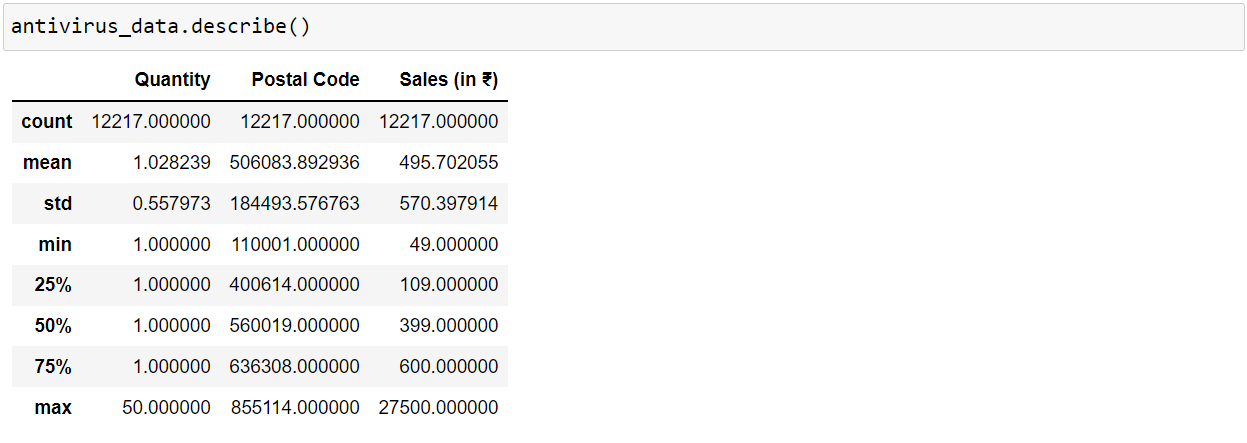


Using the ‘.value\_counts()’ method to view which State has the highest number of orders.

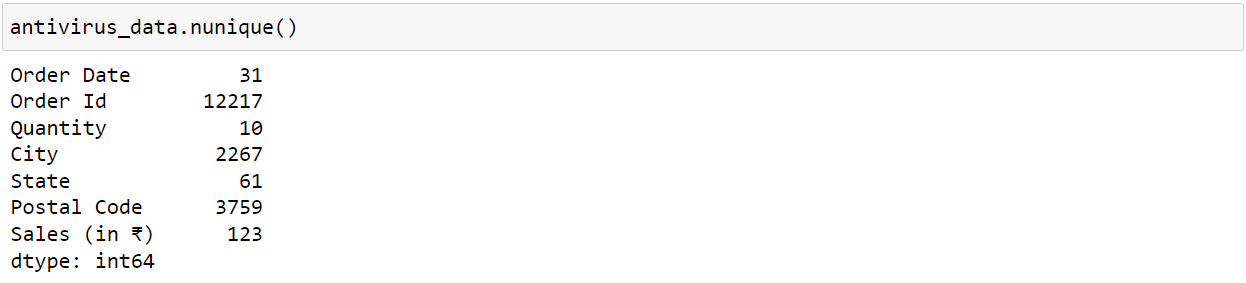


Using the ‘.value\_counts()’ method to view which City has the highest number of orders.

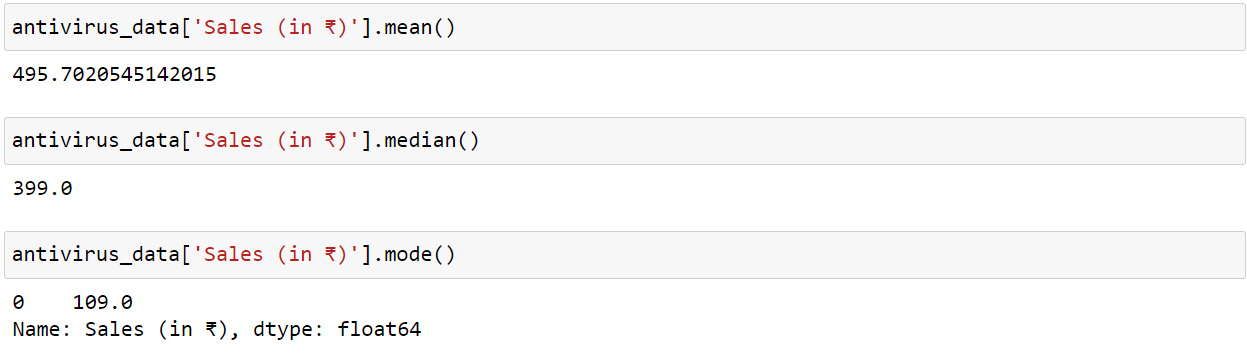




We can check the number of unique values for each column using ‘.nunique()’

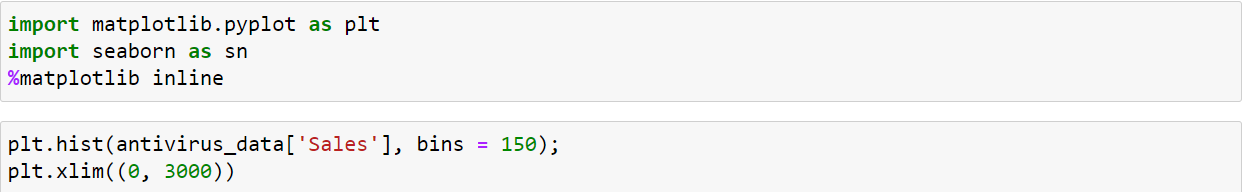


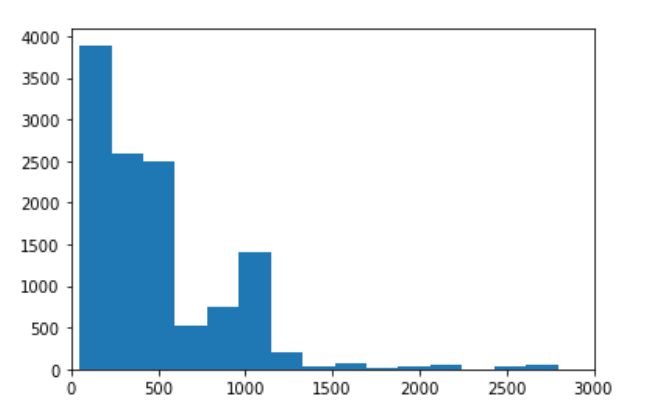
We can get the mean, median and mode for the ‘Sales’ column using ‘.mean()’, ‘.median()’ and ‘.mode()’ methods respectively



* Data Visualization

Importing the ‘matplotlib’ and ‘seaborn’ libraries for data visualisation and plotting the histogram for our measure ‘Sales’ using ‘.hist()’

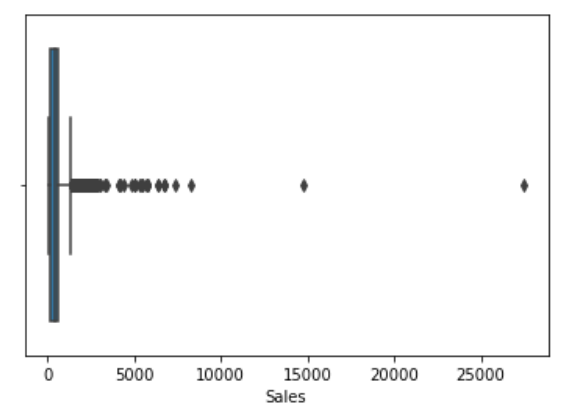




Observing the Histogram we can clearly see that the distribution is positively skewed.

We also can see that there are some outliers as well but in order to view it more clearly, we can plot a Boxplot using ‘.boxplot()’

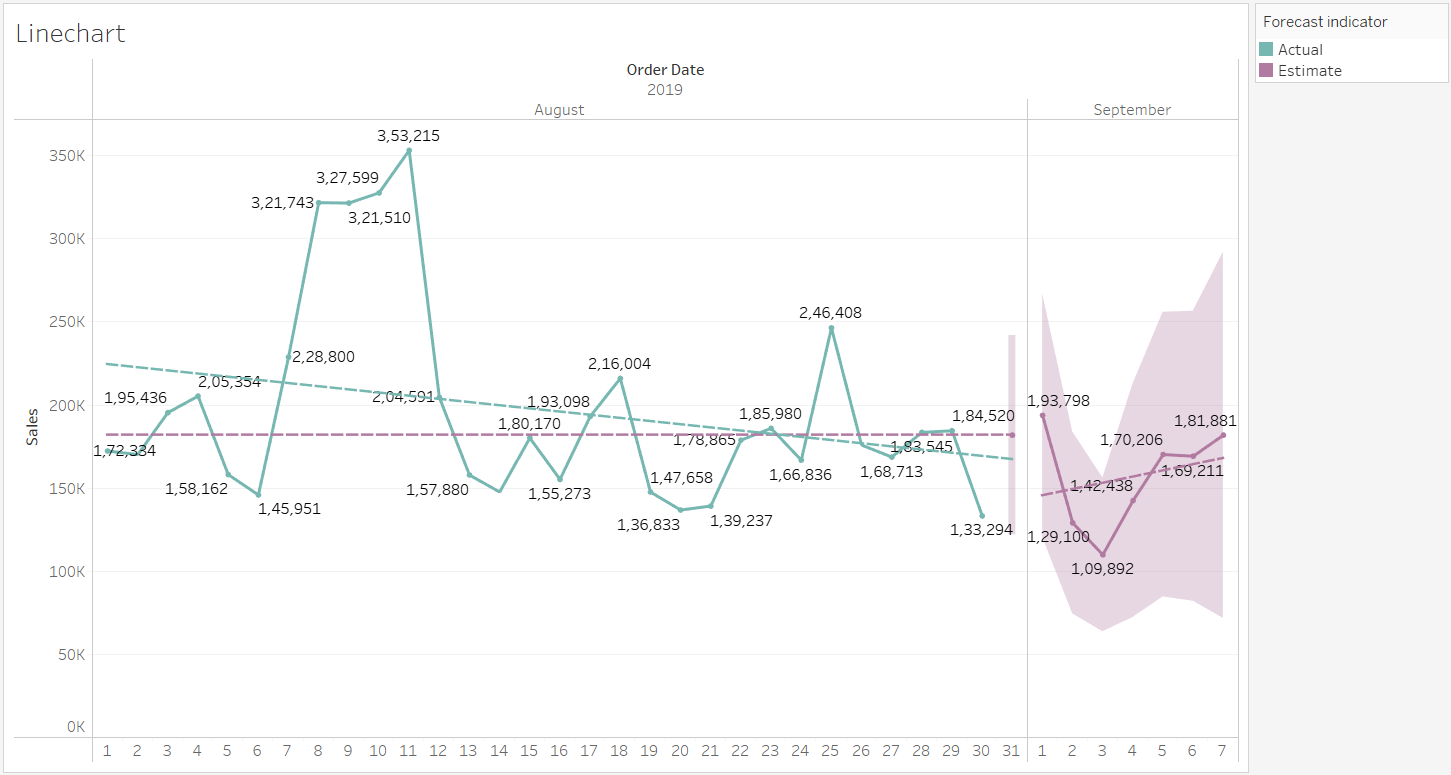




We can see that there are quite a few outliers and the majority of the data lies between 0-1000

For further visualisation, we will use Tableau as it provides ‘map’ and ‘symbol map’ charts which will help us visualise State Wise, City Wise and Postal Code Wise Sales for Kaspersky.

1. **Line chart**

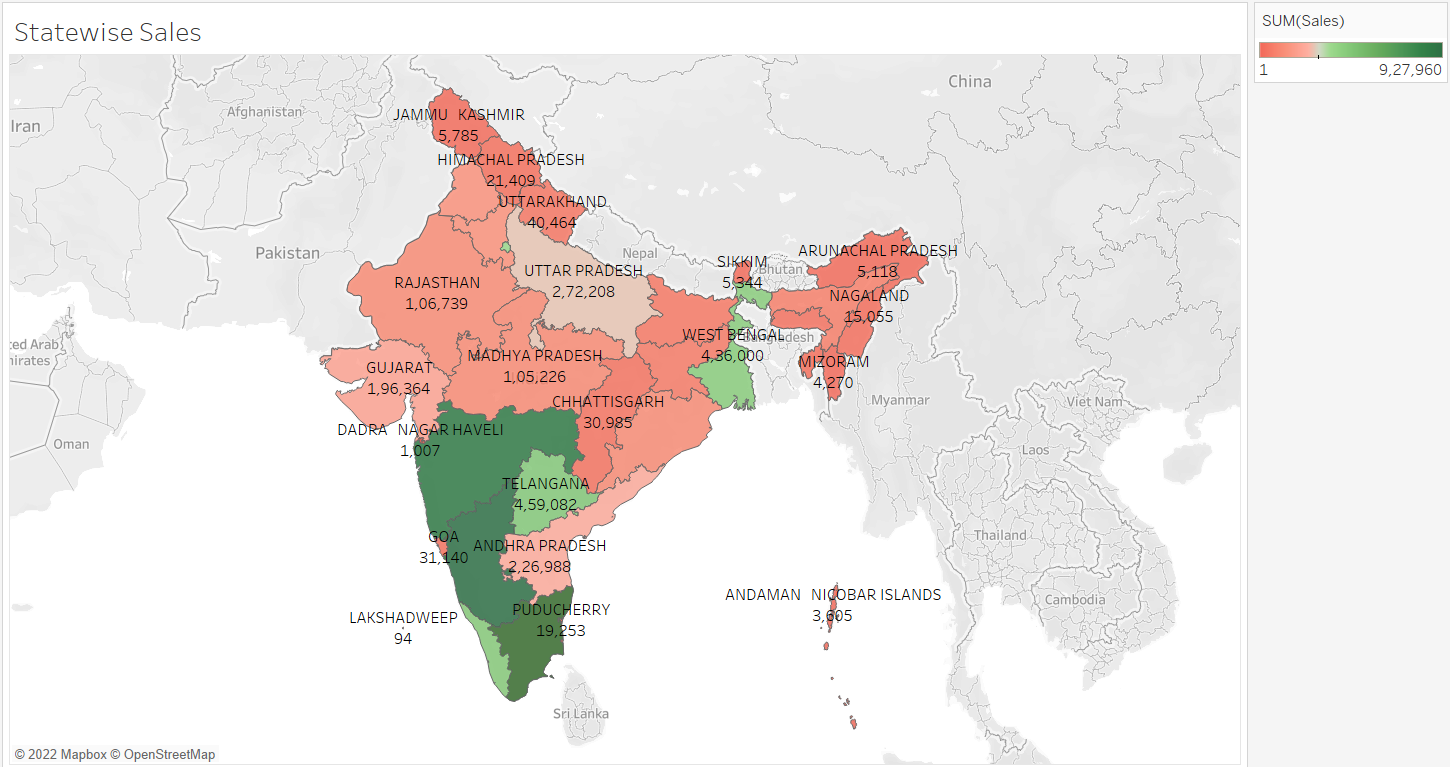


Green line represents the actual chart while the Purple line represents the estimated chart.

Plotting Sales by Date, we can see that Sales were the highest from 9th to 11th August (Rs. 3,21,743 – Rs. 3,53,215) and lowest on 30th August (Rs. 1,33,294).

Observing the trend line, we can see it has a negative slope while the forecasting feature in Tableau shows a trendline with a positive slope and predicts that the Sales for 1st of September will rise to Rs. 1,93,789 before again dropping to Rs. 1,09,892 on 3rd September and then continue to rise for the rest of the week.

1. **State Wise Sales Map**

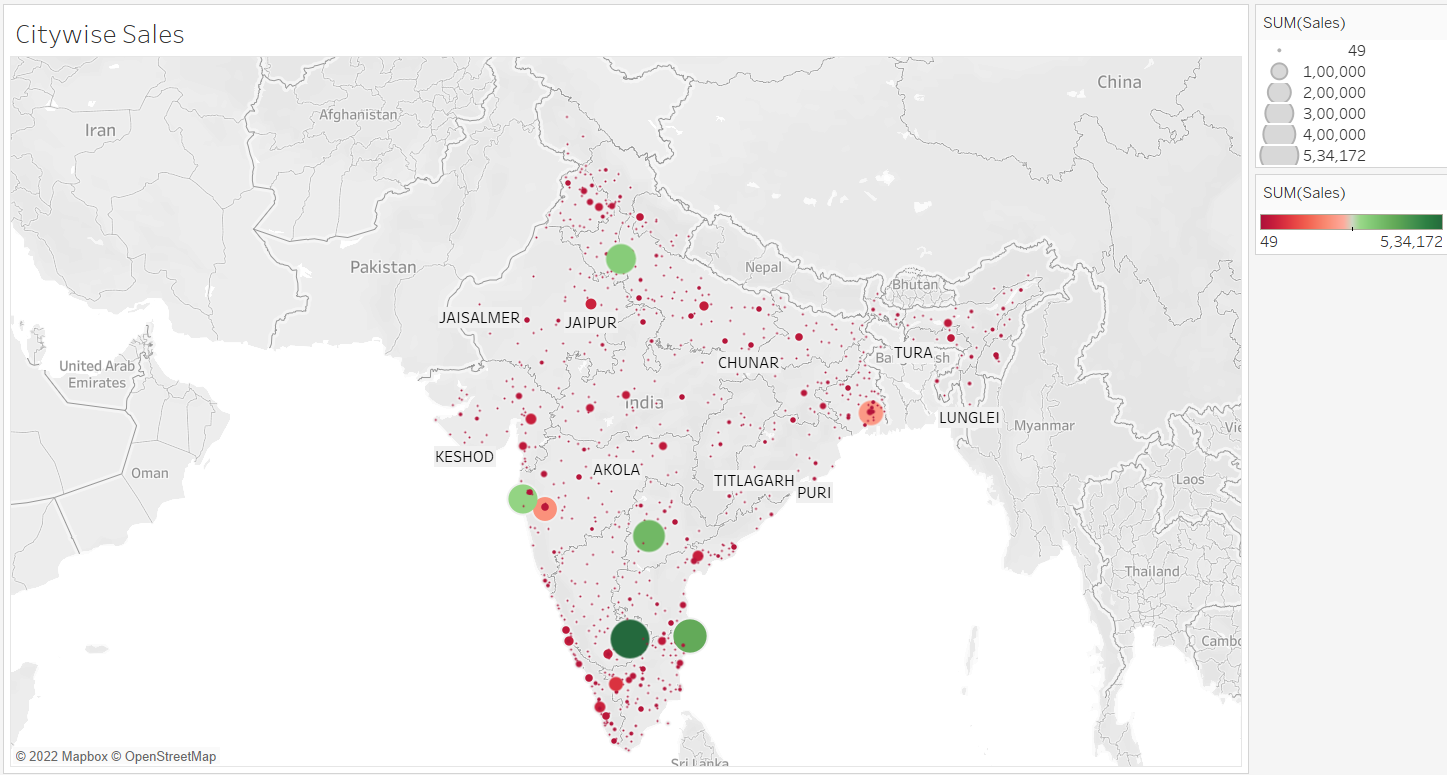


The Red-Green Diverging colour depicts the Sales in different States of India. More intense the Red colour is, the lower are the Sales for that particular State while the more intense green colour represents higher Sales.

The Top 3 states with the highest Sales are: Karnataka (Rs. 9,27,960), Maharashtra (Rs. 8,77,077) and Tamil Nadu (Rs. 8,18,901).

The Bottom 3 states with the lowest Sales are: Mizoram (Rs. 4,270), Arunachal Pradesh (Rs. 5,118), Sikkim (Rs 5,344).

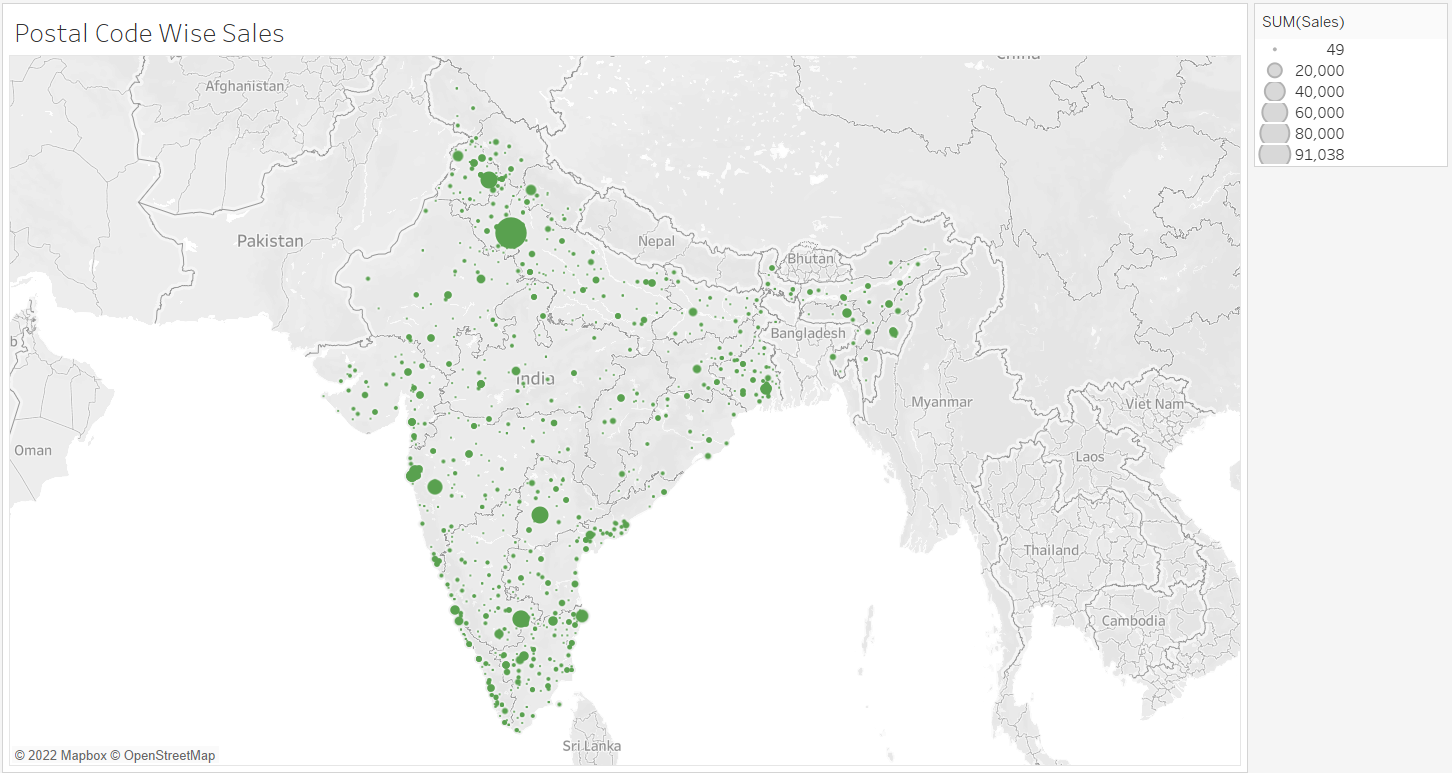
1. **City Wise Sales Symbol Map**



Here, we are using a Symbol Map wherein the size of the Bubble determines the Sales for every City. The bigger the bubble is, higher the sales is for that City while a smaller bubble represents lower Sales. To make things more clear we have also added a Red-Green Diverging here that helps us distinguish the Sales more clearly.

The Top 3 cities with the highest Sales are: Bengaluru (Rs.5,34,172), Chennai (Rs. 3,92,114) and Tamil Nadu (Rs. 3,61,635)

1. **Postal Code Wise Sales Symbol Map**



Here, we are again using Symbol Map to visualise the Sales for each PIN code.

The Biggest Bubble here is for the PIN Code 110017, New Delhi with the Sales of Rs. 91,038 which is the maximum out of any other PIN Code.

**Conclusion**

Exploratory Data Analysis helps us to look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including machine learning.

In our data, we were able to visualise the Sales by State, City and Postal Code and list down the places where the sales were the highest and where it was lowest. This would help the company to come up with different strategies to boost up Sales for those places where sales were the lowest and keep promoting Kaspersky products in the areas where the Sales are already the highest to further improve the overall growth and dominate the market.

**References**

* <https://www.bisinfotech.com/top-10-antivirus-software-of-2019/>
* <https://en.wikipedia.org/wiki/Kaspersky_Lab>
* <https://www.ibm.com/cloud/learn/exploratory-data-analysis>