Investigating Data Anomalies

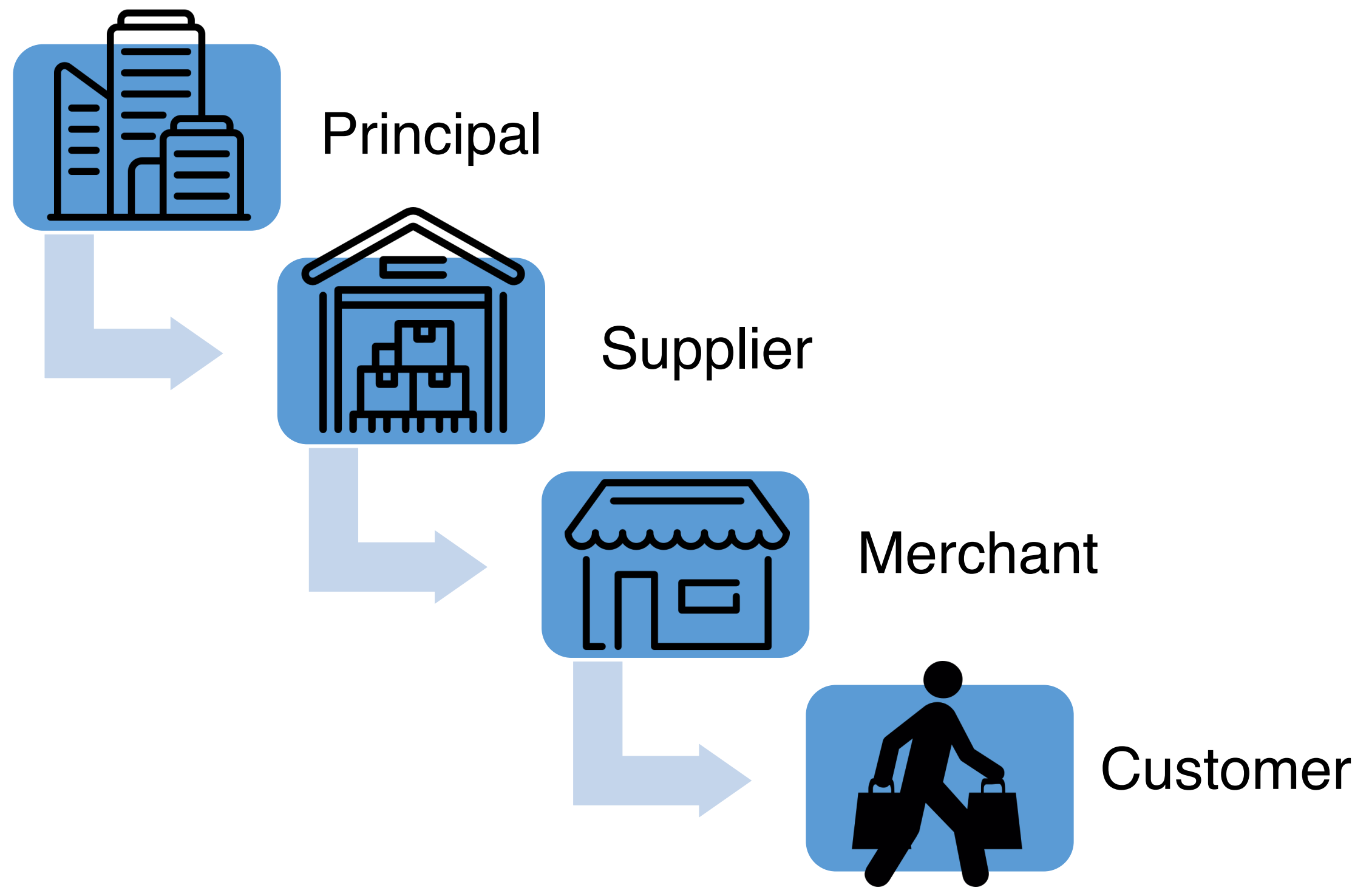
Carrie Snodgrass
Advisor: Dr. Ying Li
Department of Mathematics and Computer Science

Summary

Accurate data is critical to robust data analysis, whereas anomalies can provide a barrier to truthful and relevant observations. We conduct exploratory data analysis on a transaction dataset and detect numerous outliers. We focus on a subset of the data, comparing the prices, quantities, and amounts spent per order, before and after outliers are removed from the dataset.

Introduction and Data

In Indonesia, 82% of people buy their daily goods from small traditional merchants[1]. Traditional merchants sell to their customers in their neighborhood through face-to-face interactions, using mostly cash. These merchants make up 80% of FMCG (Fast Moving Consumer Goods) retail[1]. These merchants purchase from suppliers, who distribute from principal brands, as shown in the graphic below.


Principal
Supplier
Merchant
Customer

AwanTunai is a company that aims to digitize this supply chain. They help small merchants grow their businesses, through affordable loans and digitization[2]. The data for this project was taken from orders made through their digital system. Attributes are recorded for each order, as described in the following table.

Field	Meaning
id	index
order_id	number identifying order
placed_at	time and date of order placement
merchant_id	merchant that placed the order
sku_id	product ordered
top_cat_id	category of product
sub_cat_id	subcategory of product
qty	quantity of product ordered
price	price of one product (in Rupiah) (1 USD = 15,027 Rp)

Each of the IDs has been randomized to protect the privacy of the company and the merchants. There are 336,461 records in the data set.

Objective

The focus of this exploratory data analysis is on anomalies, patterns in data, or data points that are out of place. We focus on outliers, “observations that lie abnormal distances from other values”[3].

Where are anomalies in the data set, and what effect do they have on exploratory data analysis?

Methodology

Process

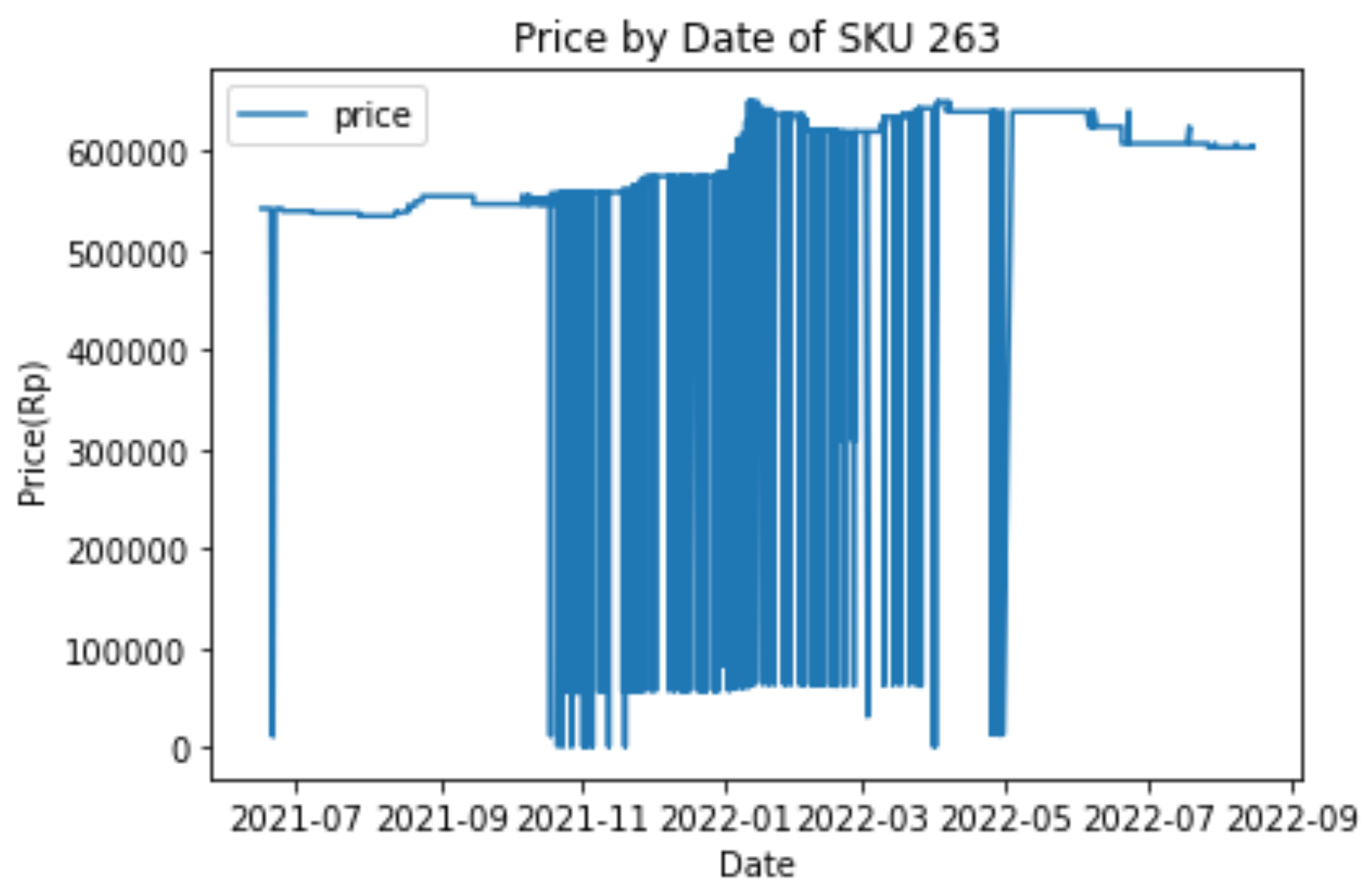
We use Python and Jupyter notebooks[4] for exploratory data analysis (EDA). The method of EDA focuses on a question for investigation. The next step is to run a query to help answer the question. Based on the query results, the question will be refined or another question, based on the original, will be asked. The process repeats as the findings and questions grow more accurate and relevant. This method of EDA allows for focus and application to specific problems.

Defining Outlier

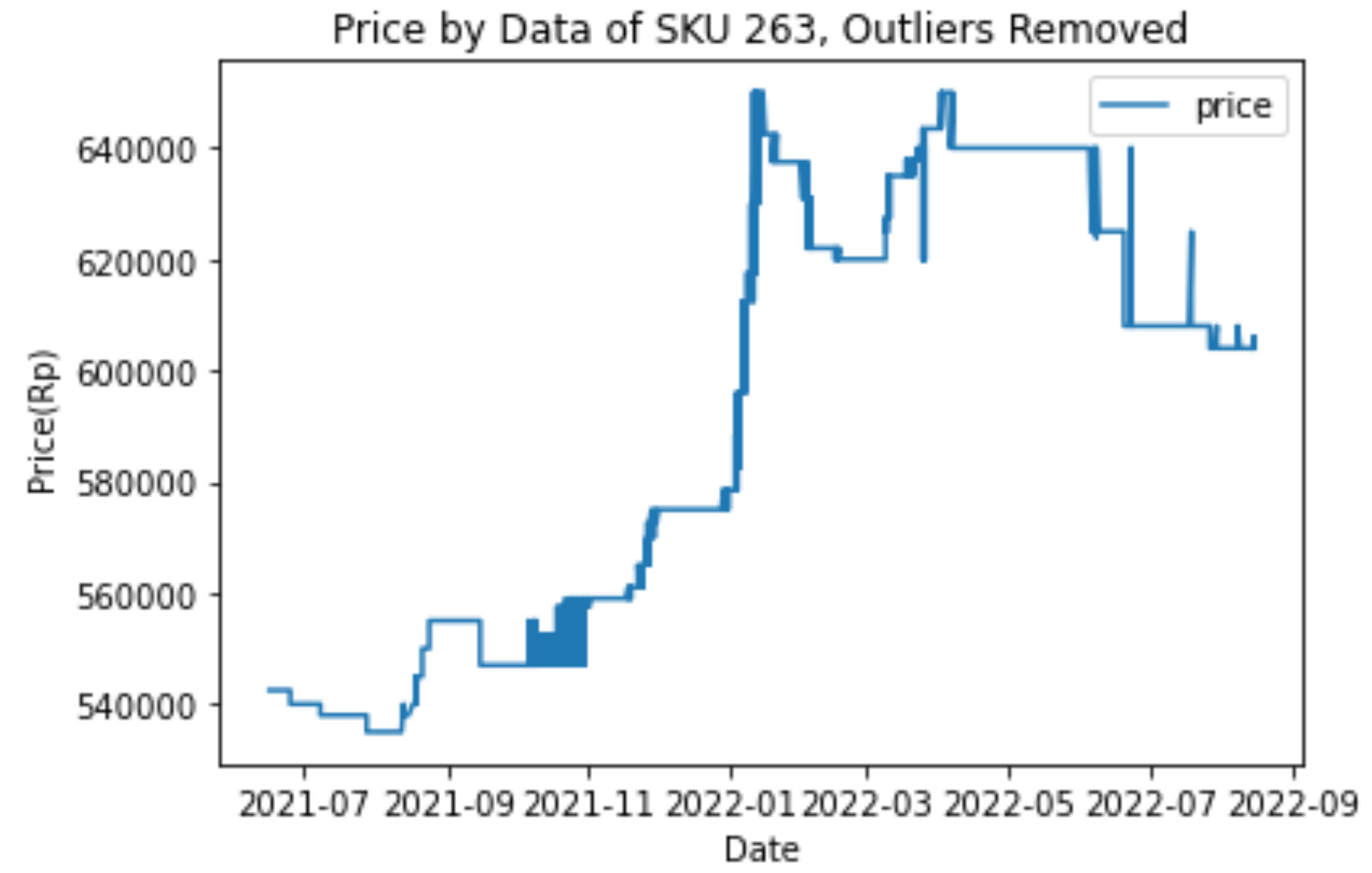
To detect outliers, we employ the concept of the interquartile range (IQR) which is defined as the distance from the 25th percentile to the 75th percentile. Any point that is outside 3 times the interquartile range above the 75th percentile and 3 times the interquartile range below the 25th percentile are considered outliers.

Results

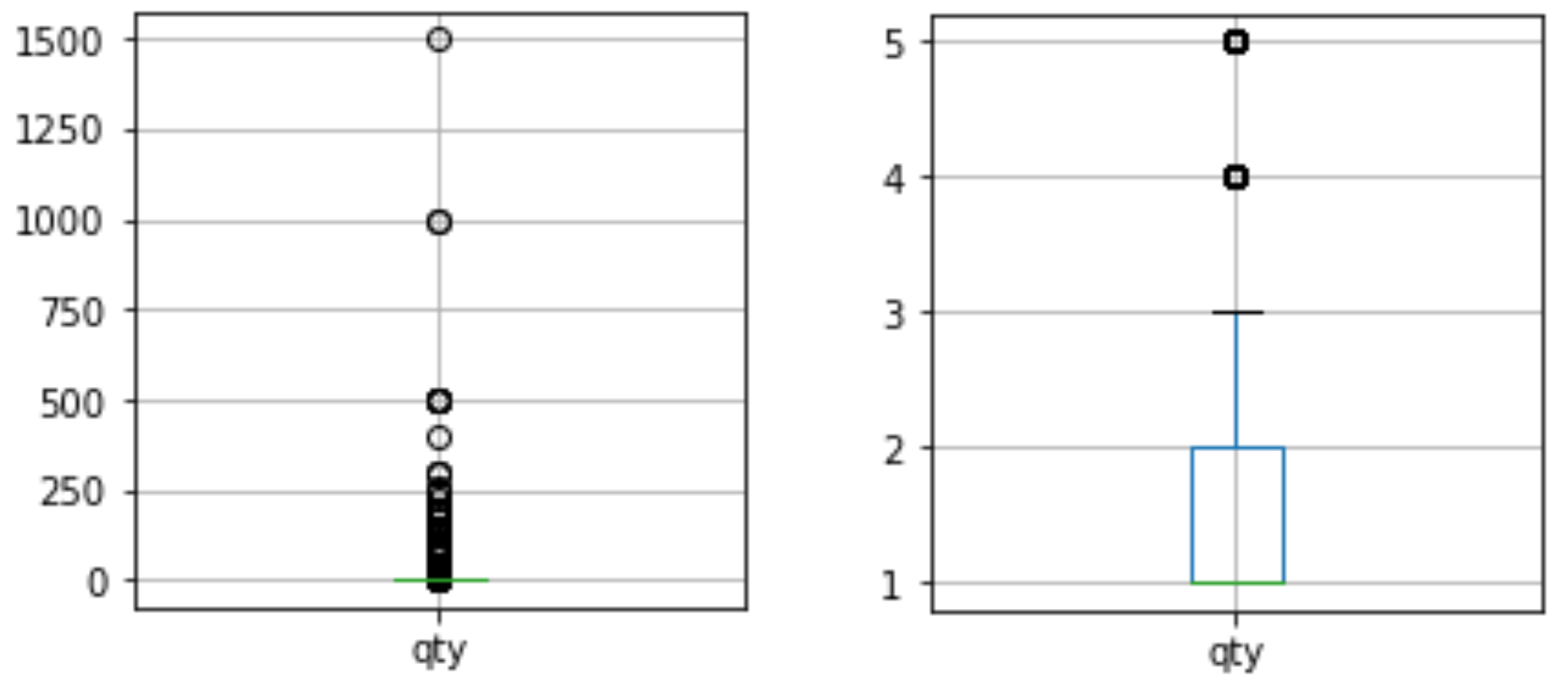
Outliers were present in the prices of each of the 10 most ordered SKUs (Stock Keeping Unit). We use the SKU with ID 263 as an example to demonstrate anomalies in the dataset. There are 4548 orders of SKU 263. The following graph contains the price against the date of the order.



The price of SKU 263 has a lower fence of 463,750. There are 155 prices that are outliers, falling below this fence. Of these, many appear to have a 0 dropped off the end of the price. The interquartile range is from 557,500 to 620,000, and 62 outliers fall between 55,750 and 62,000. This is an example of how data entry can result in significant errors.

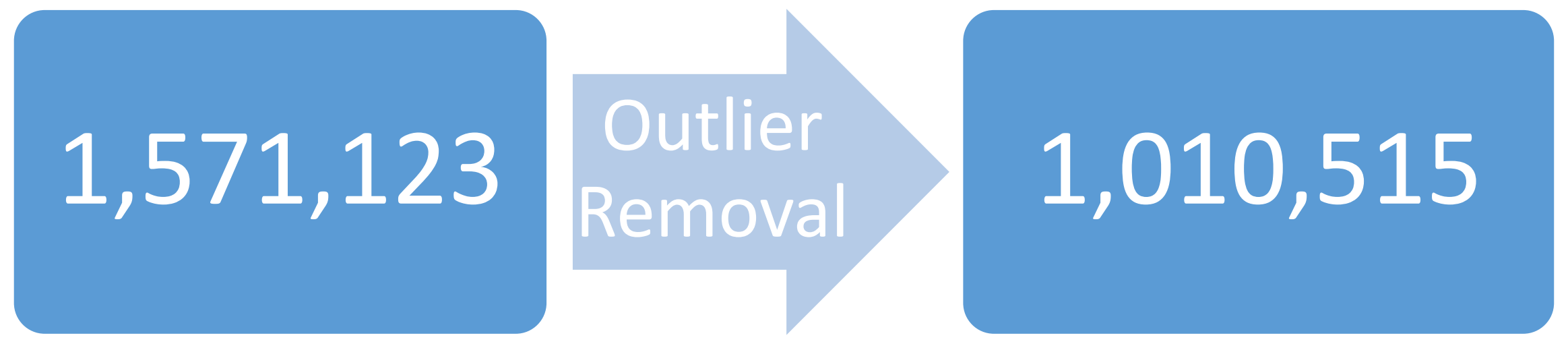


The graph of the price by date of SKU 263, with outliers removed, is shown above. The mean price of SKU 263 before the outliers are removed is 568,506 and the mean price after is 586,671, a 3% increase.



Next, the outliers are removed from the quantity of each order of SKU 263. The boxplots above describe the quantity per order of SKU 263 before (left) and after (right) the outliers are removed. 303 orders are removed.

Spent is a field created by multiplying the price and the quantity for each order. With the outliers removed from both price and quantity, the mean amount spent for orders of SKU 263 changes from 1,571,123 Rp to 1,010,515 Rp, a 35% decrease.


1,571,123
Outlier Removal
1,010,515

Average Spent per Order (Rp) Before and After Outlier Removal

In USD, this is a transition from an average of 104.56 to 67.25 dollars per transaction.

Conclusion

Outliers in data obstruct the ability to visualize data and draw conclusions. We found a 35% difference in average spent per order for SKU 263 between the average before and after removing outliers. Removing outliers can significantly change conclusions and decision-making. Sufficient analysis was not completed to recommend the removal of all outliers, and this would be an area for deeper investigation. We do recommend an investigation of systems for data recording to improve accuracy.

References

[1] Inventory Purchase Recommendation For Merchants in Traditional FMCG Retail Business, Y. Li, D. M. Robani, V. Suciu, J. He-Yueya, The 9th IEEE International Conference on Data Science and Advanced Analytics.

[2] Indonesia’s Best Supply Chain Financing Services. AwanTunai. (n.d.). Retrieved September 19, 2022.

[3] National Institute of Standards and Technology. (n.d.) *What are Outliers in the Data?*. Engineering Statistics Handbook. Retrieved September 16, 2022.

[4] The pandas development team. (2020, Feb). *Pandas-dev/pandas: Pandas.*

Acknowledgements

Thank you to AwanTunai, Wheaton College, and Wheaton College Alumni Association.