Project - 1

Analysis of Puchase dataset

Viswalahiri H | Intern (Exponentia DL) | 09/05 – 23/05



About *this* Dataset

This dataset describes purchase orders in District of Columbia in the fiscal year of 2014. More specifically, it was specified that the amount spent on each of these orders was more than $2500.

Numerically this dataset can be described as

Consisting of the parameters (**COLUMNS**)

- Purchase Order Number

- Agency

- Commodity

- Vendor Name

- Order Date

- Purchase Amount

Consisting of the 14946 entries, with multiple fields (**ROWS**)

In this analysis, we will look for optimum sales times throughout the month, ideal month for sales, effective vendors, and more insights that can help us promote or enhance sales.

Visualization in Excel

Before we start Visualizing our data in Excel, there exist many entries that have sold orders below the specified mark (2500).

Thus, let’s use those entries to keep them under the classification as small-tier vendors.

---------------------------------------------------------------------------------------------------------------------------------

**Business Objective:**

To quantify and give shape to our data in such a manner that we can extract overlooked and unrecognized information to implement decisions

---------------------------------------------------------------------------------------------------------------------------------

To find an ideal breaking or dividing point for the data, it is better to take an outlier of the data instead or the mean. It will be useful for deciding which vendors are within the limit of the outlier, and cross the outlier.

The **Larger Outlier of the data is 131281**

Thus, we can divide the vendors as

- Small Tier (< 2500 total dollar worth of P.O. sold)

- Medium Tier (<131281 and >=2500 total dollar worth of P.O. sold)

- Large Tier (>=131281 total dollar worth of P.O. sold)

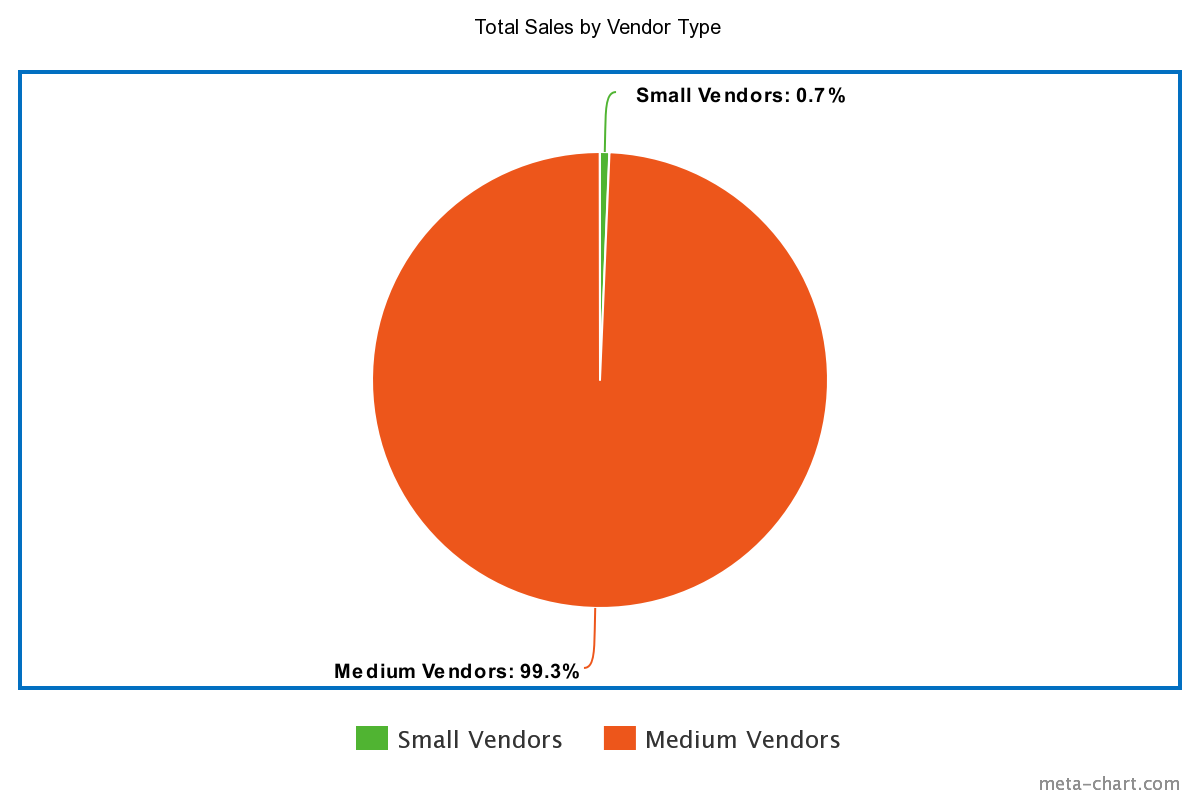
1. **Distribution of Vendors**

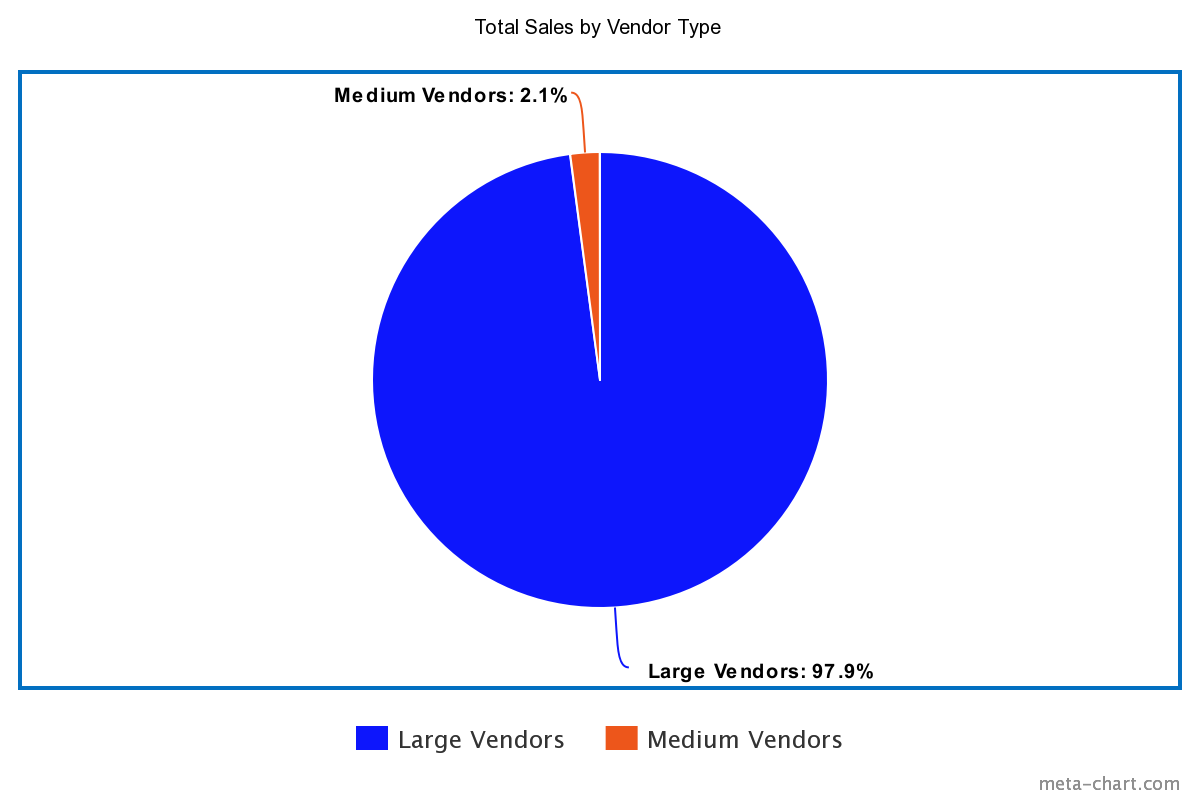
The amount of small vendors is considerably smaller than the percentage of Medium and Large vendors. Medium vendors occupy more than 50% of the total vendor count.

1. **Overall Sales of Vendors**

The sales of Medium Tier Vendors and Small Tier Vendors are extremely inferior when compared to the Large Tiered.

To get a proper understanding, let’s look at comparisons between adjacent tiers.





1. **Top Ten Vendors (in terms of total sales)**

These vendors have proven themselves to be of high importance in the market. These vendors are thus, highly known and established. Through help from the government and further marketing, global reach is highly probable to increase sales, and thus the GDP of DC.

We observe that the Company Chiaramonte-Hess Joint Venture and SKANSKA USA BUILDING deals with Construction Management Services. Whereas THE COMMUNITY PARTNERSHIP deals with Emergency Shelter.

1. **Top Ten Vendors (in terms of sale per deal)**

This measurable quantity is extremely business-oriented, as we can assess which company has more potential in sealing deals, and produce in bulk or extensively.

This information can be used by the government to promote the sale, or come to an understanding between the companies, in an attempt to supplement the GDP.

1. **Companies with the most potential**

Companies have potential, if they have a large Purchase Order Amount, and have only a few transactions.

If boosted up further, then these companies can find more clients and contribute to the total amount in a much better way.

Potential Companies include:

-Chiaramonte-Hess A Joint Ventu

-THE COMMUNITY PARTNERSHIP\HOME

-SKANSKA USA BUILDING

-DEFENSE LOGISTIC AGENCY

-COMPASS GROUP USA, INC

-ALLIED BARTON SECURITY SERVICE

-PFC ASSOCIATES LLC

-CORRECTIONS CORPORATION OF

-VERMONT ENERGY INVESTMENT CORP

These were assessed by the presence of the companies in both tables. (Total Sales, Total Sales / No. of Sales)

1. **Best Months to do Business**

Upon Quantitatively looking at the Average Sales per day in a monthly fashion, for the Medium Tier Vendors, we see that October and November are very fruitful months for The Middle Tier.

Similarly, if we looking at the Average Sales per day in a monthly fashion, for the Large Tier Vendors, we see that February and **December** are very fruitful months for The Large Tier.

However, February’s information is not reliable, since there was one transaction in February that influenced the month.

December has a larger number of successful large transactions in December.

1. **Best Weekdays to do Business**

According to the graph below, the best or most business conducive days of the week vary upon the type of Vendor sealing the deal.

Since this is across a wide data-set, we can be sure that this is a trend and not a coincidence.

Upon Quantitatively looking at the Average Sales per day in a monthly fashion, for the Medium Tier Vendors, we see that Thursday and Friday are very fruitful days.

So, government agencies can reschedule important transactions with **Medium** tender involvement to a **Thursday** or **Friday** for better results.

Upon Quantitatively looking at the Average Sales per day in a monthly fashion, for the Large Tier Vendors, we see the first three working days of the week are the most fruitful days.

However, this isn’t extremely reliable, as one deal which was equivalent to 4 other deals was made on a Tuesday.

Wednesday and Monday both are reliably fruitful for large tender deals.

However, this information isn’t completely reliable, as the presence of outliers influences the chart. Thus removing the presence of outliers, we get the following chart.

From this, we can infer that Thursday and not Friday is the best day to do business upon.

1. **Most Active Government Agencies**

The following chart explains that education is given prime importance in DC. This can be inferred as the top 2 departments are related to education.

1. **Most Successful Government Agencies in Total Sales**

Displaying Outliers

---------------------------------------------------------------------------------------------------------------------------------

**Business Objective:**

To remove extremities from our data, to get a clearer and more unbiased perspective while dealing with our data.

---------------------------------------------------------------------------------------------------------------------------------

Outliers are extremities of the data that obscure the values of the different dimensions of data that include mean, mode, median and skewness.

Outliers can be properly found with the use of IQR (Inter Quartile Range). This describes the least or maximum values that are supposed to be considered.

After removing all the values which are 0 and outliers, we obtain the remaining dataset which consists of 12693 values

For the raw dataset, if we apply the following code, we get results as listed.

Code:

Summary(raw)

Results:

|  |  |
| --- | --- |
| Min. | 0 |
| 1st Qu | 1701 |
| Median | 10000 |
| Mean | 137141 |
| 3rd Qu | 53533 |
| Max. | 85153000 |

Thus the IQR (Inter-Quartile Range) is Q3-Q1=51832

Taking the factor of IQR as 1.5, we see that the acceptable ranges of data are

Q1 – 1.5\*IQR <= Acceptable >= Q3 + 1.5\*IQR

This in turn makes the acceptable range of values as

(-76047) – (131281)

For the processed dataset, if we apply the following code, we get the results as listed

Code:

Summary(processed)

Results:

|  |  |
| --- | --- |
| Min. | 0.01 |
| 1st Qu | 1183.66 |
| Median | 7051.60 |
| Mean | 20558.35 |
| 3rd Qu | 27000.00 |
| Max. | 131257.42 |

Thus the IQR (Inter-Quartile Range) is Q3-Q1=51832

R-CODE

* raw <- read.table(file = "clipboard",sep = "\t", header=TRUE)

# Copy the data that you would like to process in Excel onto the Clip-

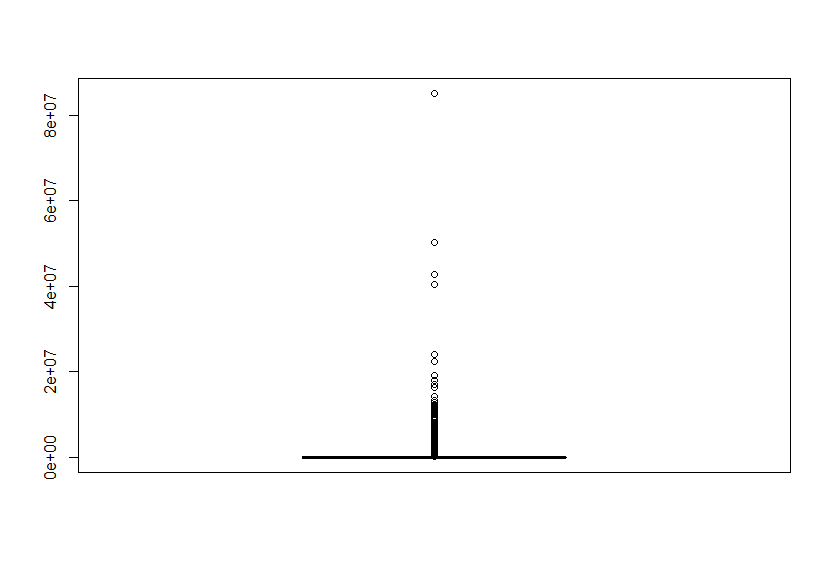
board along with the column name.

* raw<-raw[which(raw$po\_amount>0),]

# Remove all values that are equal to 0.

* boxplot(raw)

# Displays a boxplot that is obtained from the raw data.



# This box-whisker plot looks completely unfeasible to work with.

# Thus, we must remove the outliers.

* raw.numbers<-as.numeric(as.vector(as.numeric(unlist(raw))))

# # Unlist helps in coercing the list object to type ‘double’. Numeric

# helps in making the list object as a numeric type through which we

# create a vector and make this vector as a numeric type again for the

# sake of convenience.

# 

* Q1=quantile(raw.numbers,0.25)
* Q3=quantile(raw.numbers,0.75)

# #Q1 and Q3 represent the creation of Quartiles. These are quartiles

# that are based upon the raw data.

* IQR=Q3-Q1

# #Inter Quartile Range is represented by Q3-Q1

* hi=as.numeric(Q3)+1.5\*(IQR)
* low=as.numeric(Q1)-1.5\*(IQR)

# # hi and low help in the overall calculation of the range of the values

# to be considered.

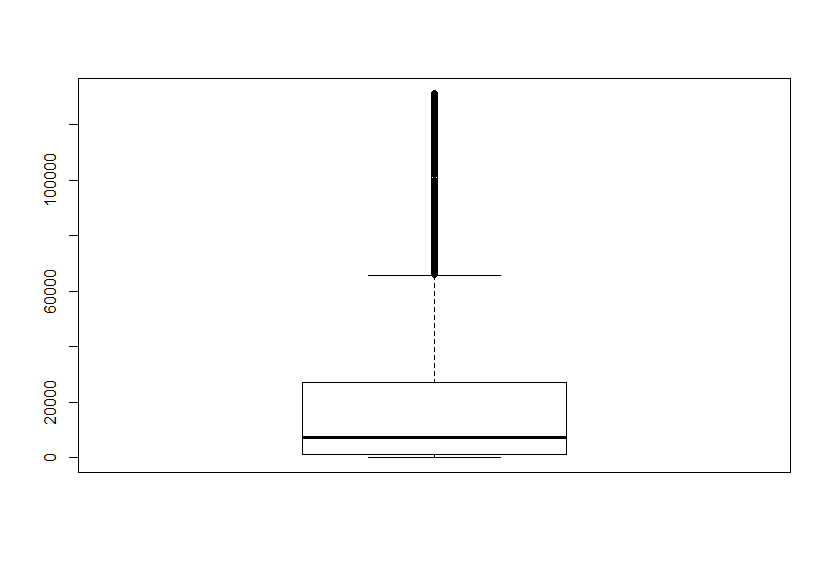
* processed<-raw[which(raw<=hi)]

# Applying conditions for the sub-setting of the dataset, that remove

the outliers from the raw data. There are no negative elements so the condition for low is unneeded.

* boxplot(processed)

# Displays a boxplot that is obtained from the raw data.



> library(tidyverse)

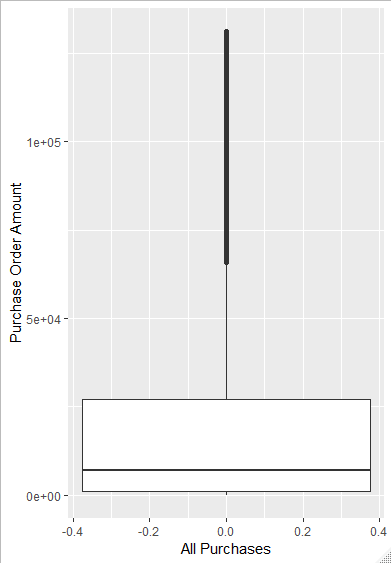
# Imports a package that is responsible for hosting the ggplot2 library

> p <- ggplot(data = process,mapping=aes(y=processed)) +geom\_boxplot()

# Stores a boxplot in the variable ‘p’.

> p+labs(y="Purchase Order Amount", x="All Purchases")

# Displays a boxplot that is obtained from the raw data.



INSIGHT

There are still many outliers that still exist however, this looks more feasible to work with.

**Initial Data:**

The boxplot of the unrefined dataset has a nearly invisible box and its upper line extends very far. This in turn implies that the data is initially extremely right skewed

We can’t judge the relationship between median and mean here as the box isn’t visible.

**Improved Data:**

The boxplot of the refined dataset has a more visible box and possesses a much larger upper line, and this concludes that the data is right sewed.

Here, we can judge the relationship between the median and mean through observation of the position of the horizontal line in the box of the box-plot.

In this box-plot **mean>median**,

Data Integrity and Sanity

Data Integrity and Sanity:

---------------------------------------------------------------------------------------------------------------------------------

**Business Objective:**

To clean our data and remove any exceptions that may lead to hindrances in our processes, or may influence the outcome.

---------------------------------------------------------------------------------------------------------------------------------

Checking For

1. Datatype mismatches:

There exist None. Column 1 and 2 have the datatypes Text and 3 and 4 have Number.

1. Variations in how values are entered:

Minimal variations exist. These don’t influence the data analysis.

1. Checking Duplicate Records and Outliers:

Outliers have been remove and there exist no duplicate copies.

Thus our data has *sanity.*

Meta-Data:

---------------------------------------------------------------------------------------------------------------------------------

**Business Objective:**

To understand the various technicalities of our data.

---------------------------------------------------------------------------------------------------------------------------------

**Initial Data Specifications:**

RangeIndex: 14946 entries, 0 to 14945

Data columns (total 6 columns):

po\_number 14946 non-null object

agency 14946 non-null object

commodity 14946 non-null object

vendor\_name 14946 non-null object

ordered\_date 14946 non-null datetime64[ns]

po\_amount 14946 non-null float64

dtypes: datetime64[ns](1), float64(1), object(5)

memory usage: 817.4+ KB

**Processed Data Specifications:**

RangeIndex: 12693 entries, 0 to 12692

Data columns (total 6 columns):

po\_number 12693 non-null object

agency 12693 non-null object

commodity 12693 non-null object

vendor\_name 12693 non-null object

ordered\_date 12693 non-null datetime64[ns]

po\_amount 12693 non-null float64

dtypes: datetime64[ns](1), float64(1), object(4)

memory usage: 595.1+ KB

Exploratory Data Analysis

( Addressed in Jupyter Notebook )

Application of Various Statistical Tests

(Addressed in Jupyter Notebook on Another Dataset)

Correlation and Covariance Checks

(Addressed in Jupyter Notebook on Another Dataset)

ANOVA

(Addressed in Jupyter Notebook on Another Dataset)

Histogram for Numerical Variables

(Addressed in Jupyter Notebook on Another Dataset)

Resultant Diagram is Depicted Below

**THE END**