Project - 1

Analysis of BFSI dataset

Viswalahiri H | Intern (Exponentia DL) | 09/05 – 21/05



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 relegate those entries and 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 the median of the data instead or the mean, as there exist outliers which greatly influence it.

The **median** of the data is **43312** (Considered from only 2500 onwards)

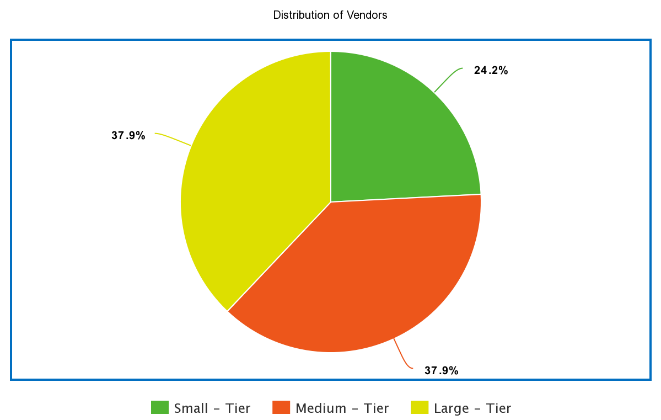
Thus, we can divide the vendors as

- Small Tier (< 2500 P.O. units)

- Medium Tier (<43312 and >=2500 P.O. units)

- Large Tier (>=43312 P.O. units)

1. **Distribution of Vendors**

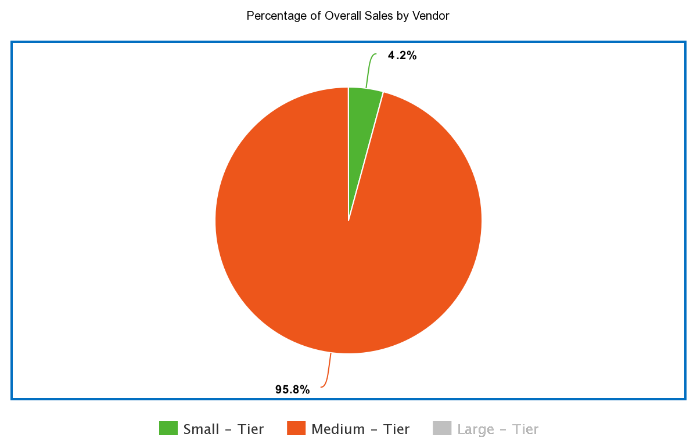
The amount of small vendors is considerably smaller than the percentage of Medium and Large vendors.****

1. **Overall Sales of Vendors**

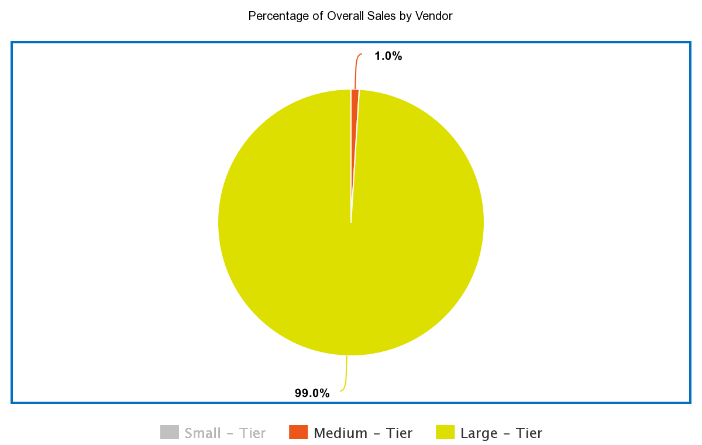
The sales of Medium Tier Vendors and Small Tier Vendors are extremely inferior when compared to the Large Tiered.

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

**Small vs Medium**



**Medium vs Large**

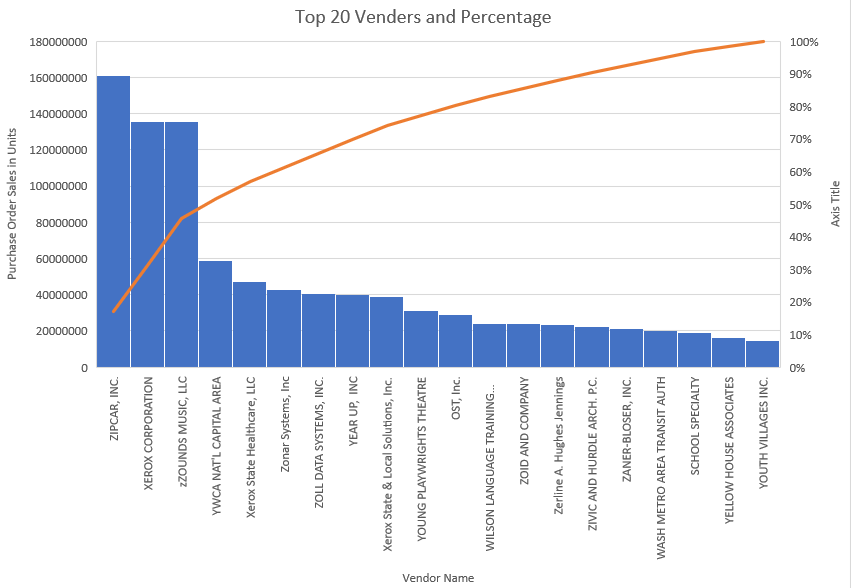
****

1. **Top Twenty Vendors and Cumulative Percentage**

**(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 ZIPCAR INC. which deals with automobile leasing, transportation and travel. Whereas XEROX CORPORATION deals with printing, leasing Xerox machines, ink and services. zZOUNDS MUSIC, LLC deals with the sale of classroom equipment.

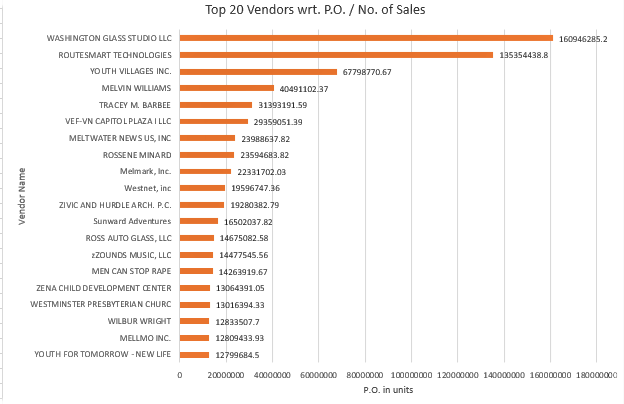


1. **Top Twenty Vendors and Cumulative Percentage**

**(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 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:

- zZOUNDS MUSIC, LLC

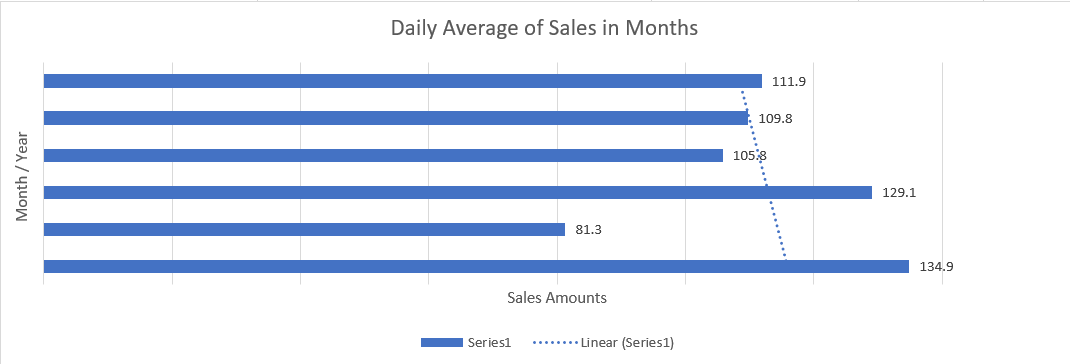
|  |
| --- |
| -ZOLL DATA SYSTEMS, INC.  -Yvonne M. Trent-Hunter  -Zerline A. Hughes Jennings |
|  |

- ZIVIC AND HURDLE ARCH. P.C.

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 Total Sales per month / Total working days, we can see the graph depicts that Business is more conducive and fruitful in the months of September and October.



1. **Best Weekdays to do Business**

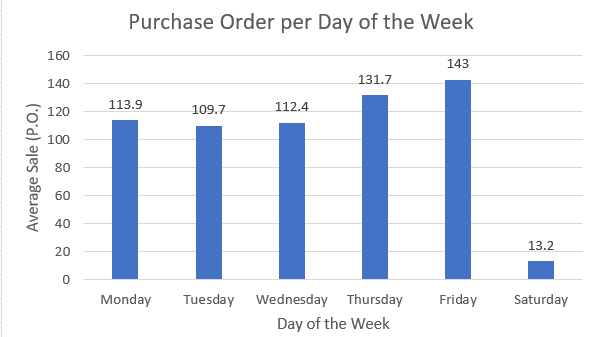
According to the graph below, the best or most business conducive days of the week are Thursday and Friday.

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

So, government agencies can reschedule important transactions with large tender involvement to a Thursday or Friday.

This can further be paired with observation 6, to place or schedule high-importance offers

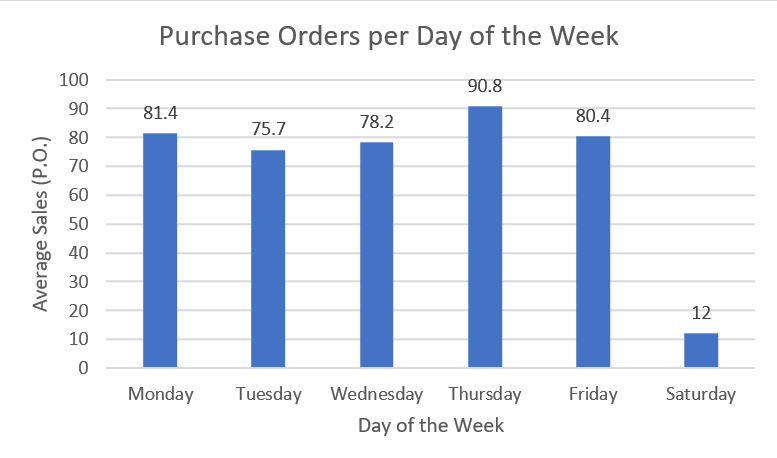
in the months of October and September.



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.

Taking the factor of IQR as 0.0278, we see that the acceptable ranges of data are

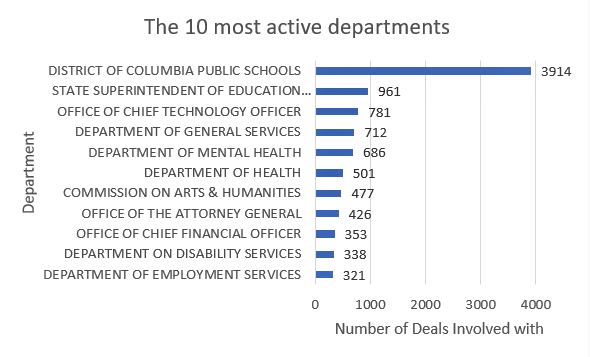
100-52997



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.



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 below 2500 and outliers, we obtain the remaining dataset which consists of 9156 values

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

Code:

Summary(raw)

Results:

po\_amount

Min. : 2500

1st Qu.: 9000

Median : 25906

Mean : 192341

3rd Qu.: 91635

Max. :85153000

Thus the IQR (Inter-Quartile Range) is Q3-Q1=82635

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

2500-215588

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

Code:

Summary(processed)

Results:

final\_amount

Min. : 2500

1st Qu.: 7680

Median : 20000

Mean : 40928

3rd Qu.: 56053

Max. : 215556

R-CODE

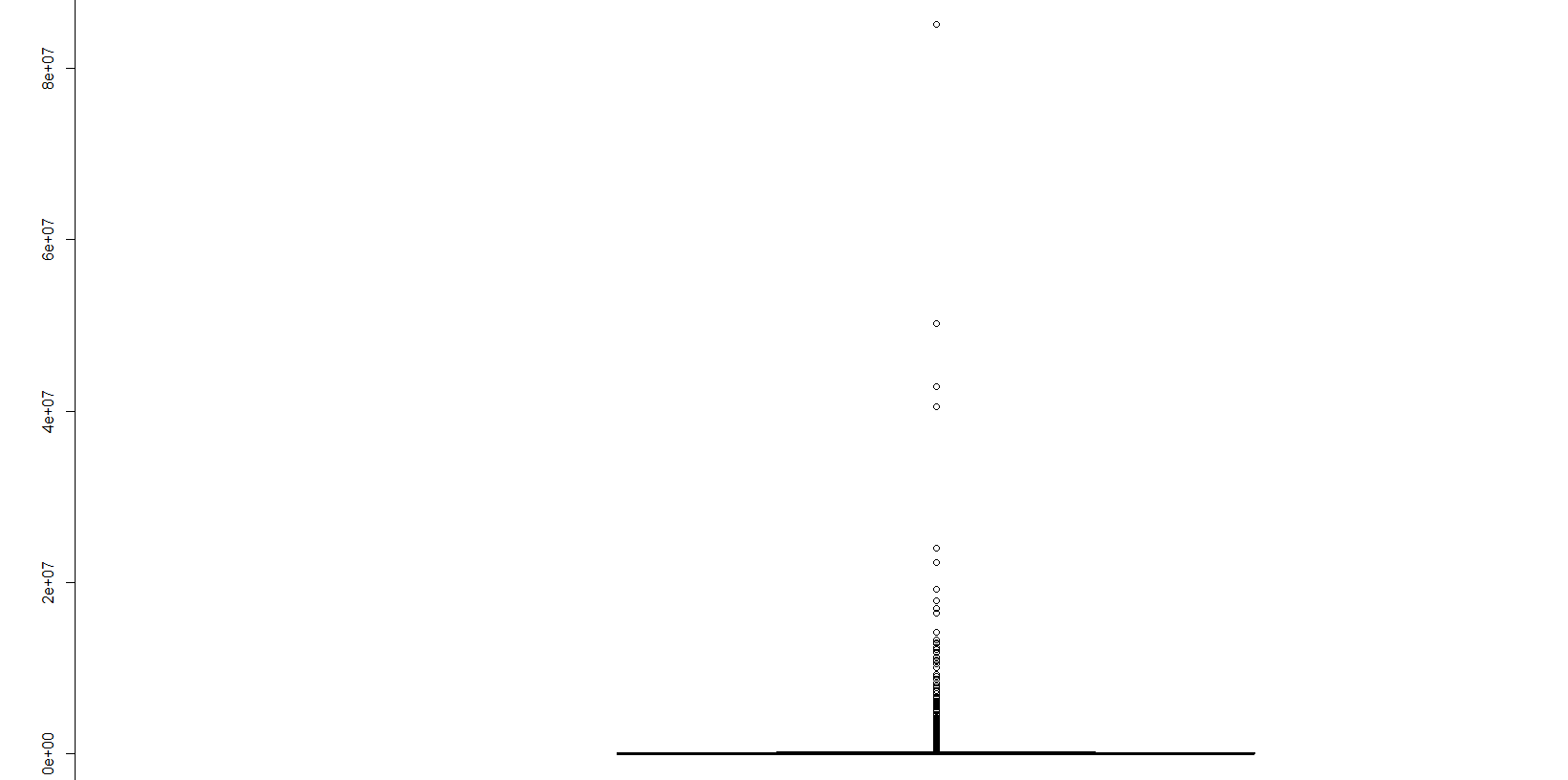
* raw <- read.table(file = "clipboard",sep = "\t", header=TRUE)

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

board along with the column name.

* 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.

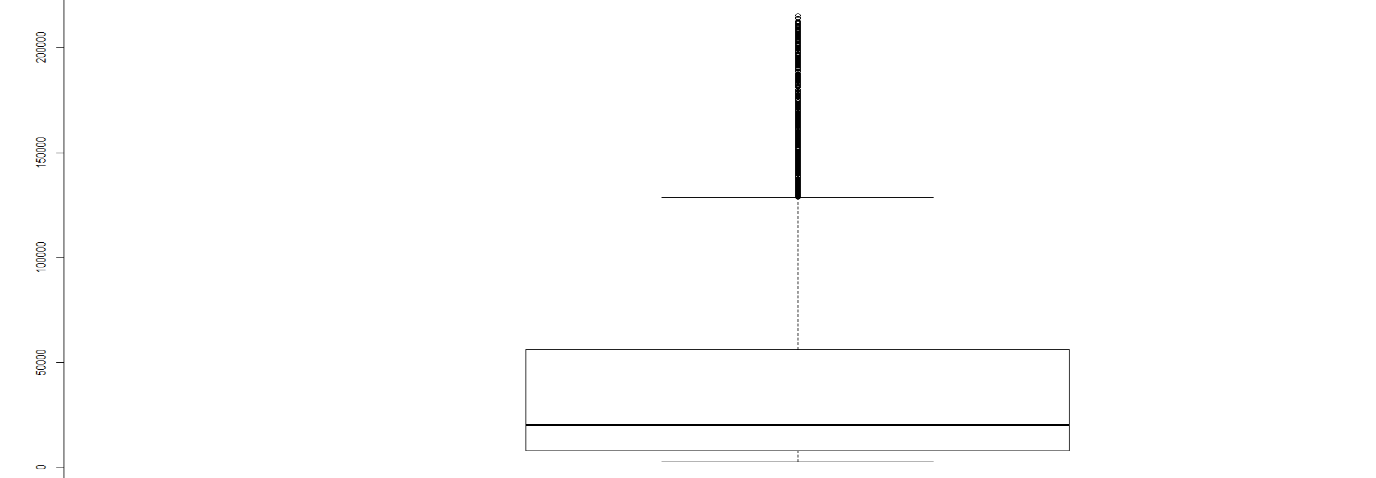
* processed<-raw.numbers[(as.numeric(raw$final\_amount)>=low) & (as.numeric(raw$final\_amount)<=hi)]

# Applying conditions for the subsetting of the dataset, that remove

the outliers from the raw data.

* Boxplot(processed)

# 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.

**Iniitial 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.

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

RangeIndex: 9156 entries, 0 to 9155

Data columns (total 6 columns):

po\_number 9156 non-null object

agency 9156 non-null object

commodity 9156 non-null object

vendor\_name 9156 non-null object

ordered\_date 9156 non-null datetime64[ns]

po\_amount 9156 non-null float64

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

memory usage: 429.3+ 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**