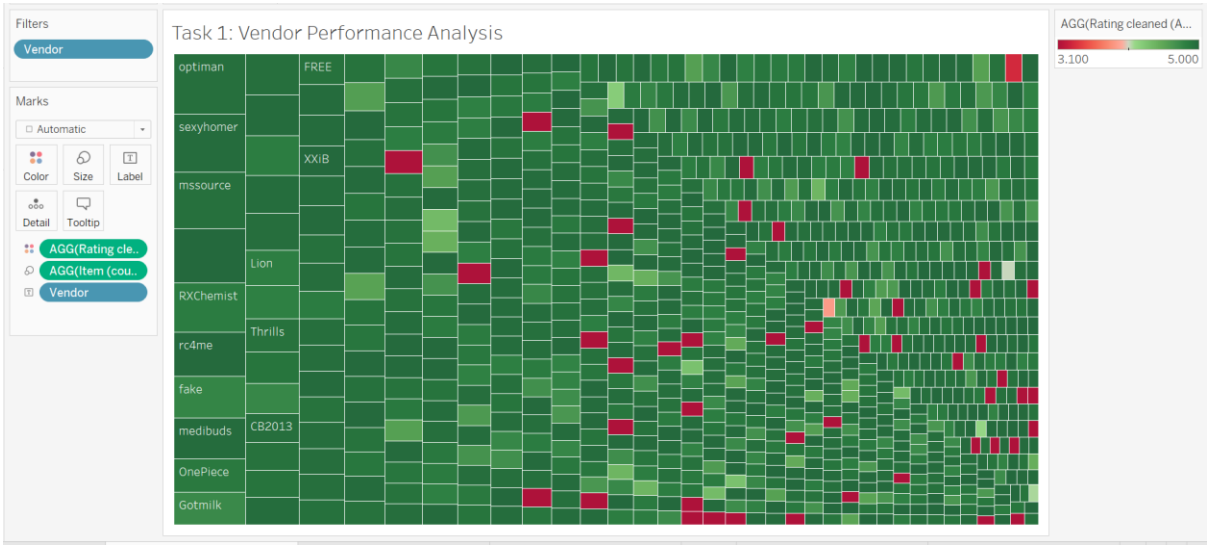
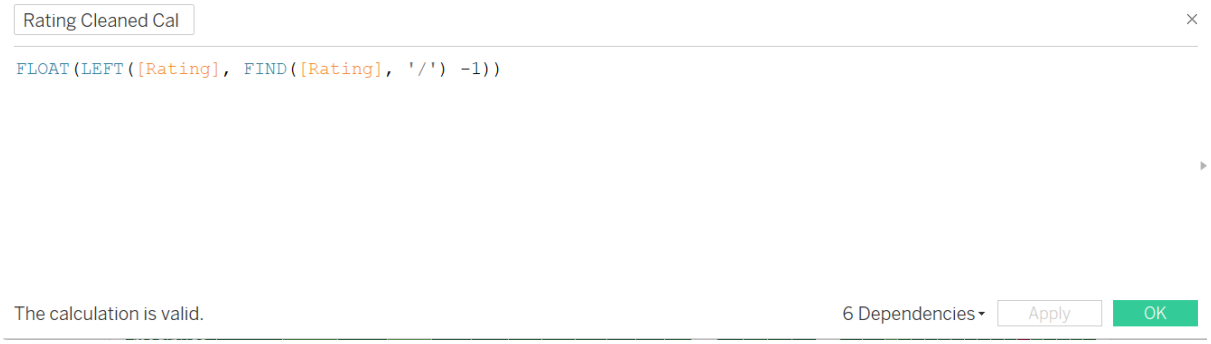
Dataset used Agora dataset: (Agora Dataset Anon n.d.)

Vendor Performance Analysis

Visualization:



Calculations screenshots:



Insights:

The above tree map gives an insight into average ratings for the top 1000 vendors. The vendor's feature is placed in the filters section and selecting only the top 1000 vendors based on total count of the items.

The ratings field has values including 4.96/5. A calculation field is generated which will extract 4.96 from 4.96/5. This is called data cleaning. This extracts the value properly from the field.

Further justification on the calculation field: The calculation field uses the rating field, finds for the position of the string /'. Once the position is found it subtracts the position by 1 and takes the left values till that number (position -1).

The map explanation:

The map gives a broader view of the number of total items of each top 1000 vendors. It shows the highest and lowest number of items along with ratings. The map also shows the highest and lowest average ratings for each vendor.

The vendor with highest count of items is **Optiman** with 885 total items.

The highest average rating is from vendor **Profesorhouse** with a rating of 5.00. Along with vendor Profesorhouse there are few more vendors who have the highest rating of 5.00.

On the other hand, the lowest rating has some issues. The lowest average rating is null. This is because the data is having issues with the values. There are some values with data such as [0 deals] or ~4/5 which need to be filtered and extracted correctly. This issue has been addressed correctly in task 2.

Questions:

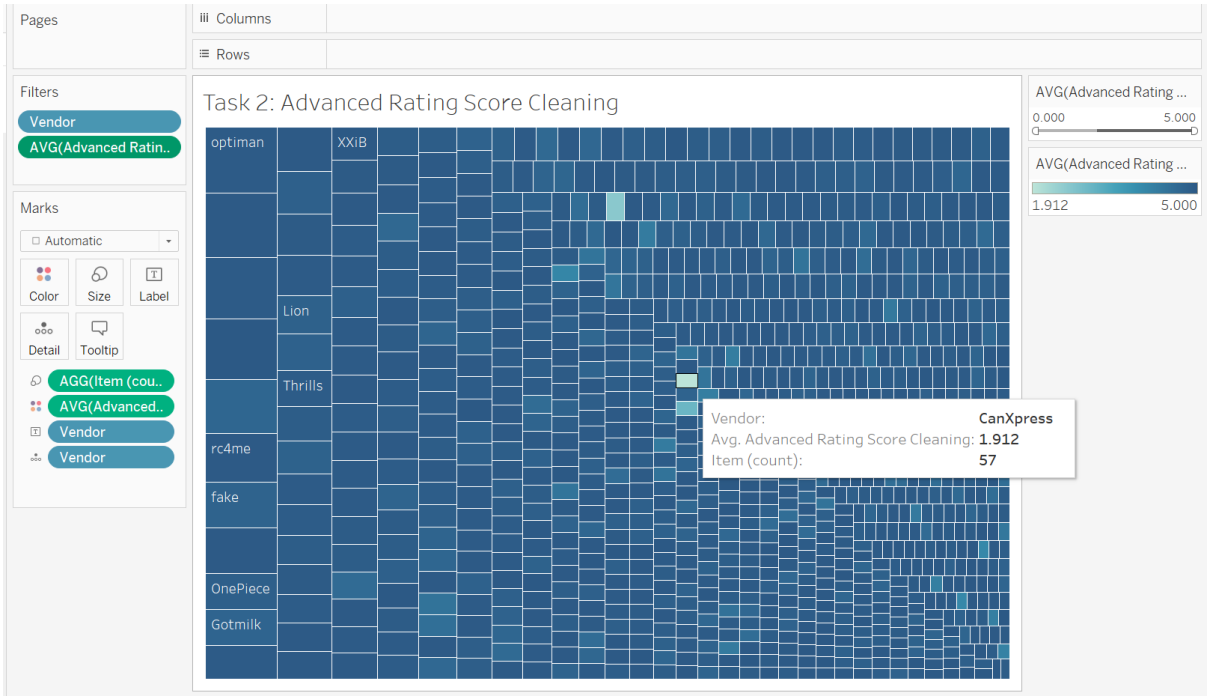
- Which vendor has the highest item count?
  - The **Optiman** vendor has the highest item count of around 885.

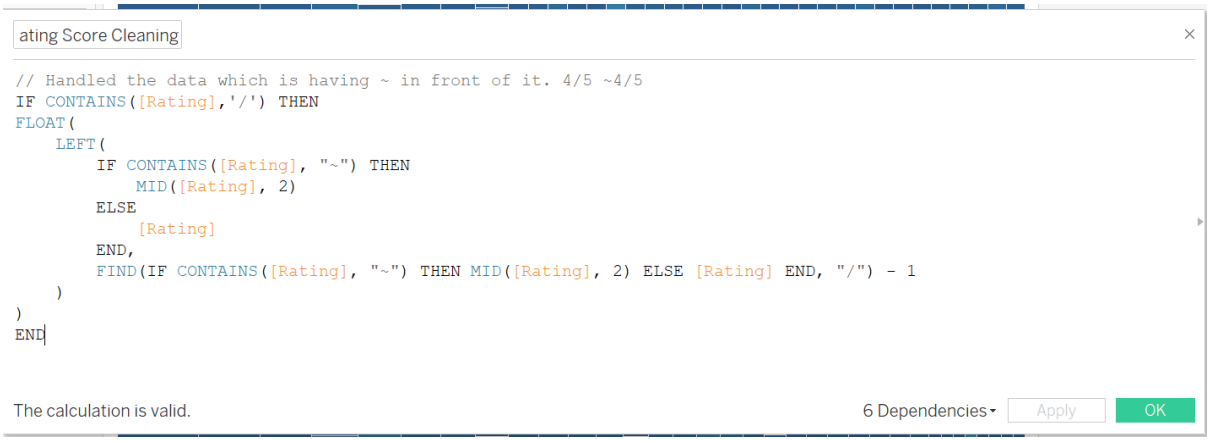
- Which vendor ratings have the lowest values?
  - The **DarkShop**, **italianconnection** and **46** more vendors have the lowest values. The DarkShop and EbookShop have the lowest ratings with the lowest and with the count of items as 33 and 185 respectively.
    For further view, there is a sheet named: **Revision sheet to check the values.** This sheet consists of the lowest values for the top 1000 vendors with their item count.

- What issues do you find for the lowest values?
  - The lowest values are null. This is due to data issues. The issues such as the ratings field have values such as [0 deals], ~4/5 which need to be filtered out and extracted properly.

Advanced Rating Score Cleaning

Visualization:





Calculations:

Insights:

The above tree map gives a highlight of the values which were causing issues in task 1 such as [0 deals], 4/5 and ~4/5.

This graph gives the average lowest value of ratings properly. This is due to the filtering and handling of the data with a calculation shown in question 2.

New questions can be formed, and solutions can be seen.

Questions:

- Which vendors have the lowest average ratings? What did they sell?
  - Vendors such as **CanExpress**, **topnotchpills** and **New_demension** have the lowest ratings of 1.9, 2.5 and 3.0 respectively. The three mentioned vendors sold **Drugs** as their main category.

Previously, in task 1 it was very difficult to get the lowest values for the vendors because of data issues. It was null value for each lowest value. This was because the data was not handled properly. After addressing the issue in task 2, the vendors with lowest value are now seen properly which also gives an insight that those vendors mostly sold drugs as their main category.

- Explain how you deal with special characters and null values using Tableau calculations.

➢ Extracted 4/5 from ~4/5, 4 from 4/5 and [0 deals] as a null value. Below is the detailed calculation with a screenshot of calculation provided above.

➢ **Calculations insights:**

This calculation handles the data which is having ~4/5 and 4/5 along with [0 deals]. First it checks it checks if the rating has /'. If yes, then it goes in the calculation for 4/5 or ~4/5. If rating does not have / then it returns null i.e it handles [0 deals] as null.

Now if ratings have / then it checks if ratings have '~'. If yes, then extracts the value from 2nd position/character which is 4/5 from ~4/5. If ratings do not have '~' then it returns ratings which means it will have 4/5 by default.

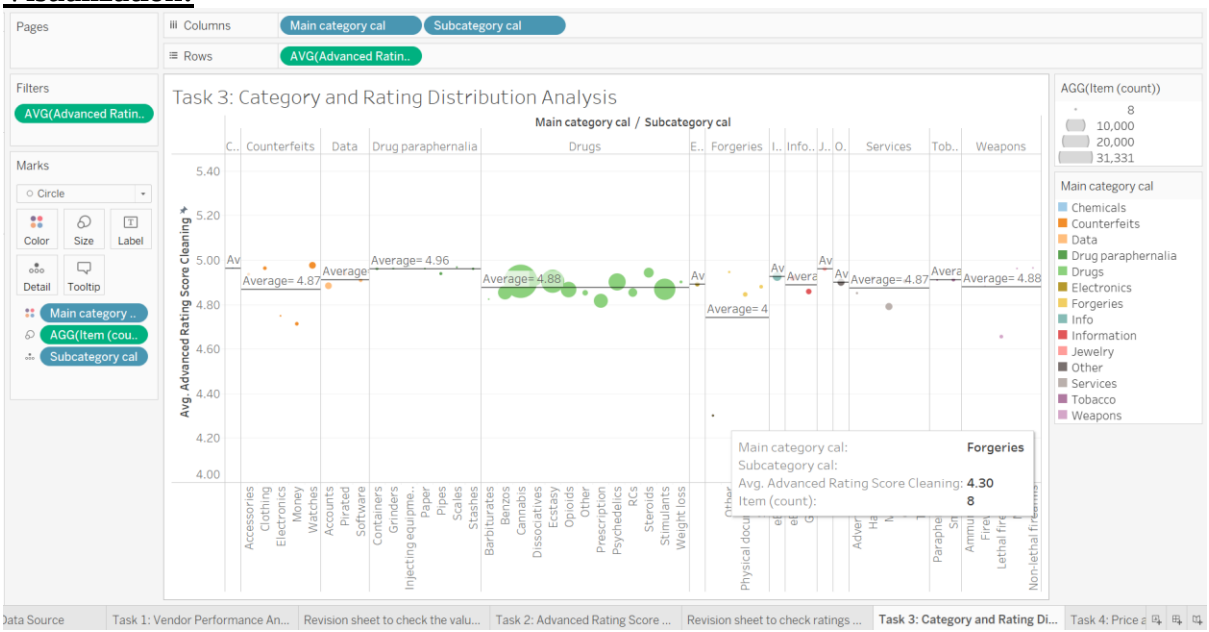
Once the value is extracted, it will find the rating if ~ then it will take the value from position 2 else it will keep the data as it is and then position will be extracted which has /. That position will be subtracted by 1 and the left function will be operated. The left function will extract the left values before \.

After performing the calculation, the calculation field is provided in the filter, and only non-null values are taken into account.

The sheet in tableau named: **Revision sheet to check ratings after performing cleaning** shows the before cleaning and after cleaning results.

Category and Rating Distribution Analysis

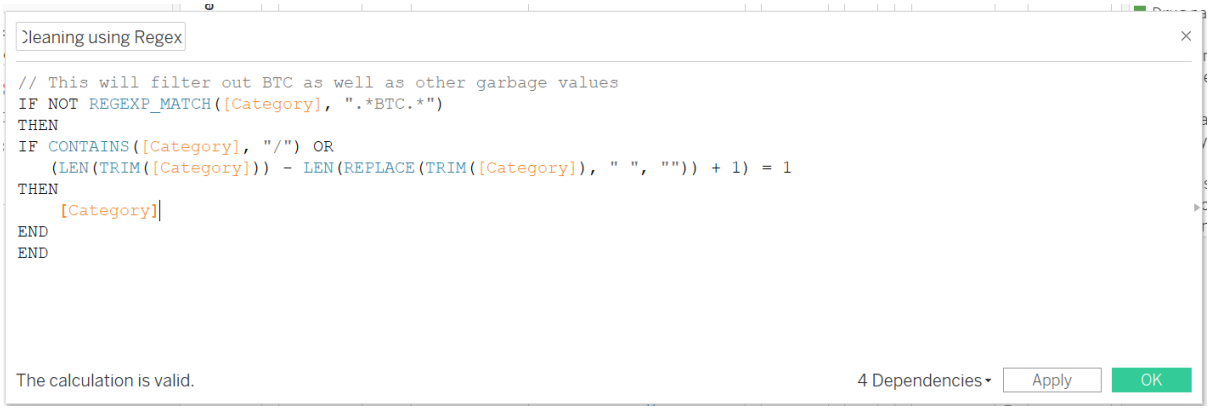
Visualization:



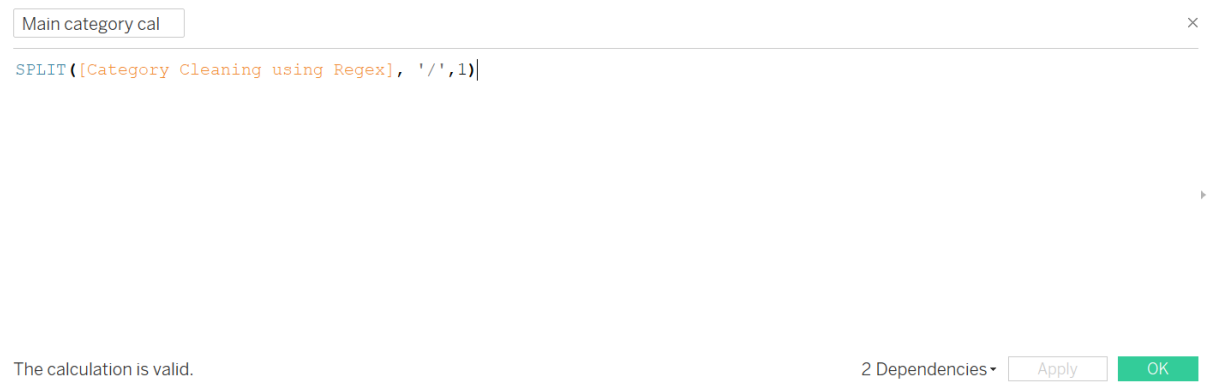
Insights:

The above graph gives the highlight of the sales based on categories with respect to their ratings.
It shows that there are few categories which have ratings below average along with their count of items. To achieve this plot, a calculation field is created which extracts the data from categories as main categories and subcategories. This distinguishing is done by /.

The screenshot explains the cleaning of the category field. It filters out the BTC as well as other garbage values.

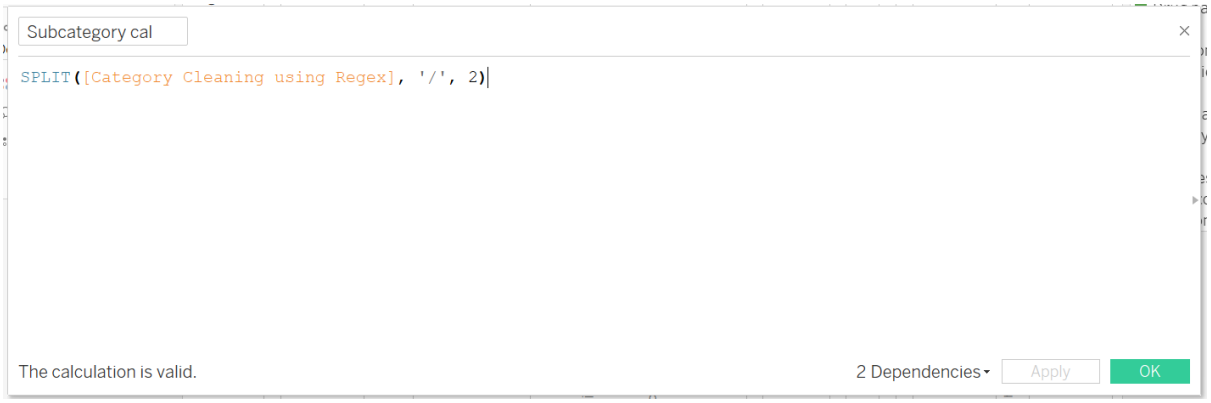


After performing the cleaning this field is used to segregate the main category and subcategory.



A split method is used to split the cleaned category field by / and the first occurrence is extracted.

Similarly, for the subcategory field calculation. The second occurrence of the cleaned category field is extracted.



Questions:

- The main category **Forgeries** has the rating lower than average of around 4.30.
- The main category **Counterfeits** with subcategory of **Money** has the rating lower than average of around 4.71.
- The main category **Counterfeits** with subcategory of **Electronics** has the rating lower than average of around 4.45.
- The main category **Services** with subcategory of **Money** has the rating lower than average of around 4.79. There are few more which can be observed from the plot.

- Despite of having low item count from Weapons, counterfeits and forgeries main categories, their rating score seems to be the highest one. This means that despite of less sales the ratings seemed to be the highest one.

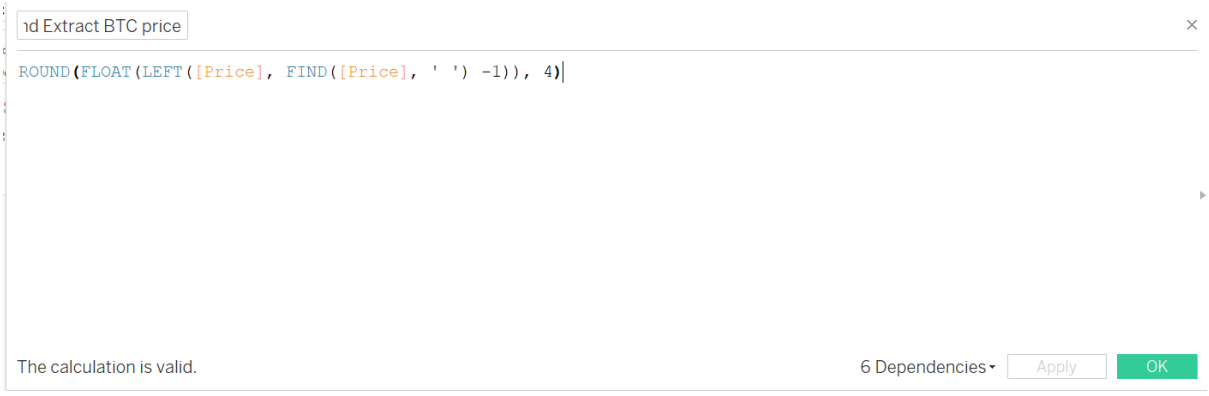
Price and Rating Correlation Analysis

Visualizations:

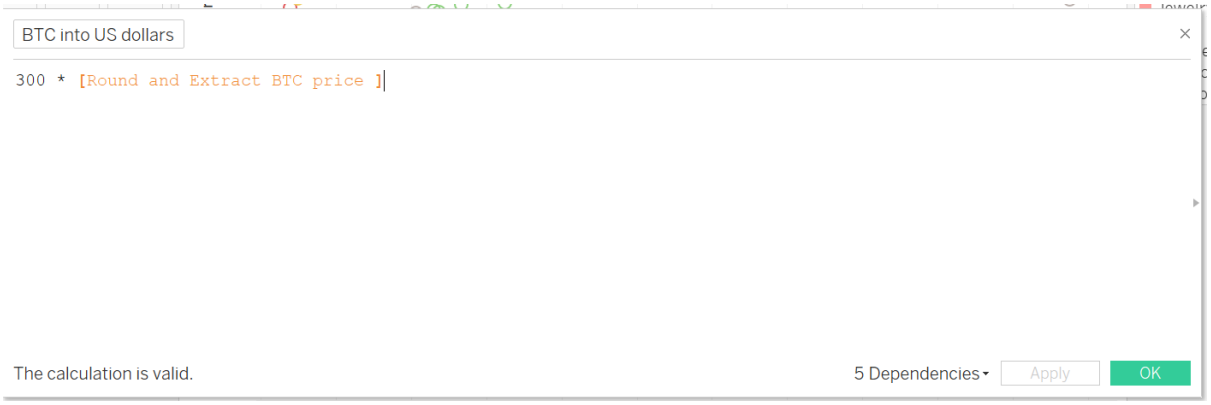
<u>Insights:</u>

The above plot gives an overview of how the average rating is related to average price with corresponding main and subcategories.
To achieve this calculation fields are created.

The price field in the dataset consists of prices in BTC. This BTC needs to be filtered out. For this a calculation field is created which will filter out BTC from the value. For example: 0.1524 from 0.152419585 BTC



The above calculation finds the space and returns the position. This position number is then subtracted by 1 and the left function is applied on price. The left function will extract the value not including the BTC. The value **0.152419585** will be extracted and then a round method will round it to 4 decimal places. The final extracted value will be **0.1524.**



The extracted value will then be multiplied by 300 to convert to USD from BTC. (1 BTC = 300 USD)

Calculations for excluding extreme values: Values >= 2000 are excluded.

Questions:

- Average ratings is inversely proportional to average price with respect to categories.
  Even though the ratings are higher above 4.8 the average price seems to be low for the categories mentioned. This can be observed closely through the plot. However, there are few categories which have higher ratings along with higher prices. These are:

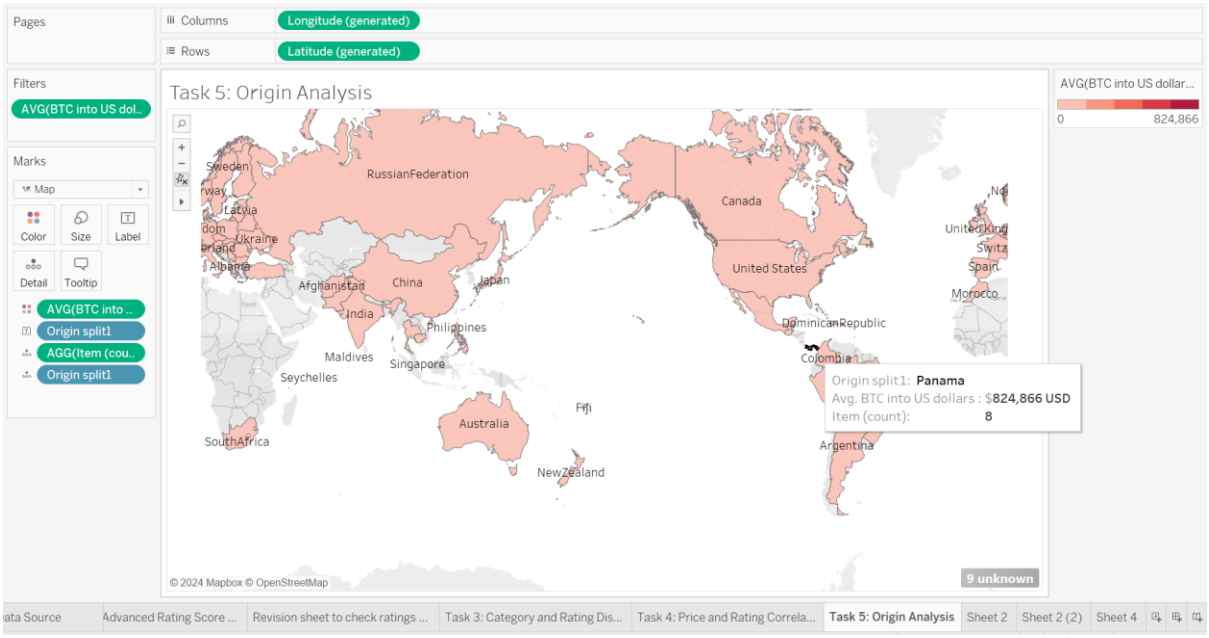
  Main category: Services
  Subcategory: Travel
  Avg ratings: 4.88
  Avg Price: 1075$

  Main category: Weapons
  Subcategory: Lethal firearms
  Avg ratings: 4.69
  Avg Price: 928$

- Categories are mostly scattered in the area where average ratings are above 4.8 and average price is less than 400 USD. These results are for the prices less than 2000.
  Moreover, there is a category of forgeries which has the lowest average ratings of 4.30 with an average price of 124 USD.

Origin Analysis

Visualizations:



Insights:

The above chart gives an idea of distribution of items in each country along with the average price. The map shows that **Panama** is having the highest average price of around 824,866 USD with count of items as 8.

To achieve this plot a calculation is needed to segregate the origin column properly into each country.

Calculations:

Here, the segregation is done on origin field, and its first occurrence is considered. This is achieved by using a split method. Here, if the origin field does not contain BTC and img (garbage value) then split the origin field and get the first occurrence. To make a note: only the first split is considered, and the data is then segregated into respective countries.

Origin split1

```
IF CONTAINS([split1], 'USA') OR CONTAINS([split1], 'US') OR CONTAINS([split1], 'United') OR CONTAINS([split1], 'America') OR CONTAINS([split1], 'U.S.A') OR
"United States"

ELSEIF CONTAINS([split1], 'Ukraine') THEN
"Ukraine"

ELSEIF CONTAINS([split1], 'UK') OR CONTAINS([split1], 'Uk') OR CONTAINS([split1], 'UnitedKingdom') OR CONTAINS([split1], 'UntiedKingdom') OR CONTAINS([split
"United Kingdom"

ELSEIF CONTAINS([split1], 'Canada') OR CONTAINS([split1], 'CANADA') OR CONTAINS([split1], 'canada') OR CONTAINS([split1], 'Canadagb') OR CONTAINS([split1],
"Canada"

ELSEIF CONTAINS([split1], 'Germany') OR CONTAINS([split1], 'German') OR CONTAINS([split1], 'GERMANY') OR CONTAINS([split1], 'GermanyGermany') THEN
"Germany"

ELSEIF CONTAINS([split1], 'Australia') OR CONTAINS([split1], 'Australianew') OR CONTAINS([split1], 'New Zealand') OR CONTAINS([split1], 'Christmas Island')
"Australia"

ELSEIF CONTAINS([split1], 'Netherlands') OR CONTAINS([split1], 'Netherland') OR CONTAINS([split1], 'TheNetherlands') OR CONTAINS([split1], 'NetherlandsAntil
"Netherlands"

ELSEIF CONTAINS([split1], 'China') OR CONTAINS([split1], 'Hong Kong') OR CONTAINS([split1], 'Hong') OR CONTAINS([split1], 'china') THEN
"China"

ELSEIF  CONTAINS([split1],'Sweden') THEN "Sweden"

ELSEIF  CONTAINS([split1],'Mexico') THEN "Mexico"

ELSEIF  CONTAINS([split1],'Latvia') THEN "Latvia"

ELSEIF CONTAINS([split1], 'EU') THEN "Europe"

ELSEIF CONTAINS([split1], 'Agora') OR CONTAINS([split1], 'World') OR CONTAINS([split1], 'Cheqdropz') OR CONTAINS([split1], 'ChristmasIsland') OR CONTAINS([s
NULL

ELSE [split1]
END
```

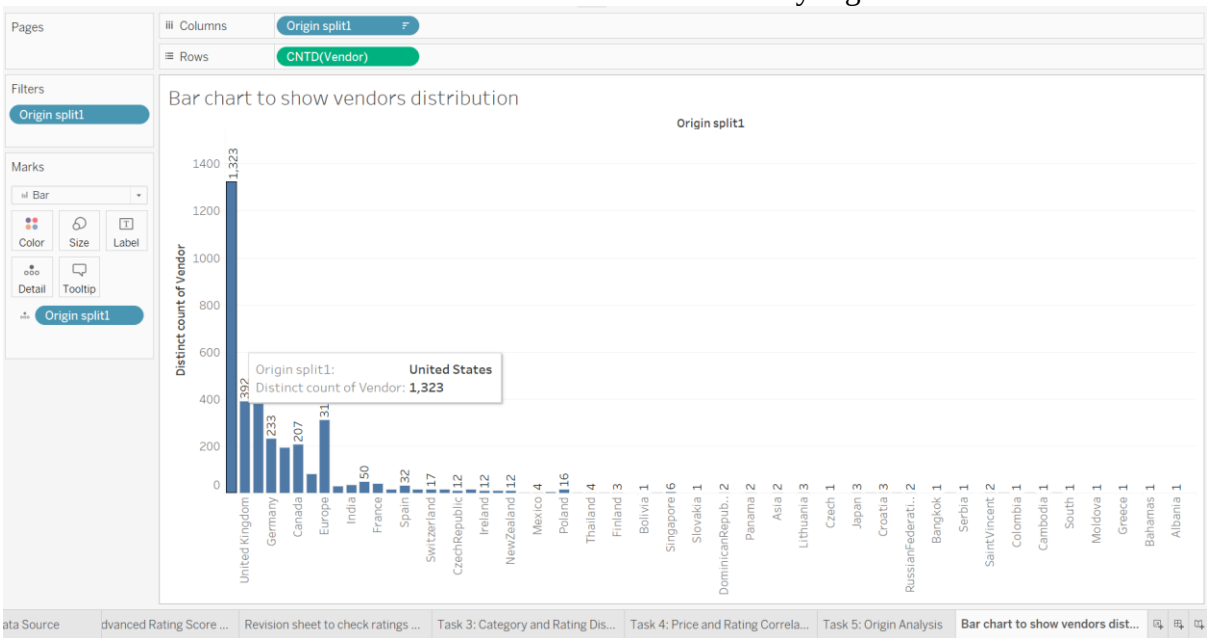
In the above calculation, data in the split1 calculation is considered. It checks if the split1 contains the strings provided if yes then that data is mapped to required country.

The null is mapped to garbage data or the data which is not related to a country. The final **origin split1** is used to plot the map.

For better reasoning the sheet named: **Sheet for the splitting analysis** is added to the tableau book which will give a good insight of filtering the origin column.

Questions:

- Use a bar chart to show the vendors' distribution. Are there any regions that dominate?



- This plot shows vendors' distribution with respect to origin. From this graph, about 1323 number of vendors are distributed from United States. Followed by United Kingdom which is of about 392 vendors. So, the regions that mostly dominate the vendors are United States and United Kingdom.

- Which origin has the highest average item price and how these compare across regions? Then, filter out extreme prices >= $2,000 USD to observe the changes.

➢ Before changing: **Panama** has the highest average item price of around 824866 USD.

➢ After filtering out the extreme price >= $2000 USD, **Mexico** has the highest average price of 1788 USD followed by **United Kingdom** with 1735 USD.