# Contents



# Index of Figures

# Week 7 Project Submission: In-Depth EDA – Team 108

## 1. In-Depth EDA

In order to perform the In-Depth EDA applied to the data received for the project case chosen with the Eficacia company, we initially take a general overlook of the data contained on each of the csv files received from the customer, and try to identify possible relations that could potentially exist between all of them, this as an initial approach in order to identify next steps such as relations, possible scenarios and which of the tables contain a higher volume of information.

For this step we left as evidence the general overview of the table sizes and number of rows, below find the evidence of this initial exploration to the data:

| File | Size (KB) | Number of Columns | Number of Rows |
|---|---|---|---|
| Agotados.csv | 88553 | 11 | 1'364.563 |
| Exhibiciones.csv | 8456 | 11 | 87.357 |
| Inventarios.csv | 2728 | 6 | 62.166 |
| Precios.csv | 2095 | 8 | 32.287 |
| Productos.csv | 17 | 7 | 233 |
| PuntosVenta.csv | 241 | 19 | 1.702 |
| Ventas.csv | 95767 | 8 | 1'397.331 |

*Figure 1 Overview of Tables information*

We can see on the pie chart below that most of the data is contained in two tables “Ventas” and “Agotados” on this initial exploratory analysis this does not give us a lot of insights about the data contained on the table nonetheless it gives us an idea of where the exploratory analysis should be focused in this case the correlation between the sales and the out of stock will be very important therefore having the higher volume of information on “Ventas” and “Agotados” could be a good indicator about the data need for the analysis and future models to be implemented

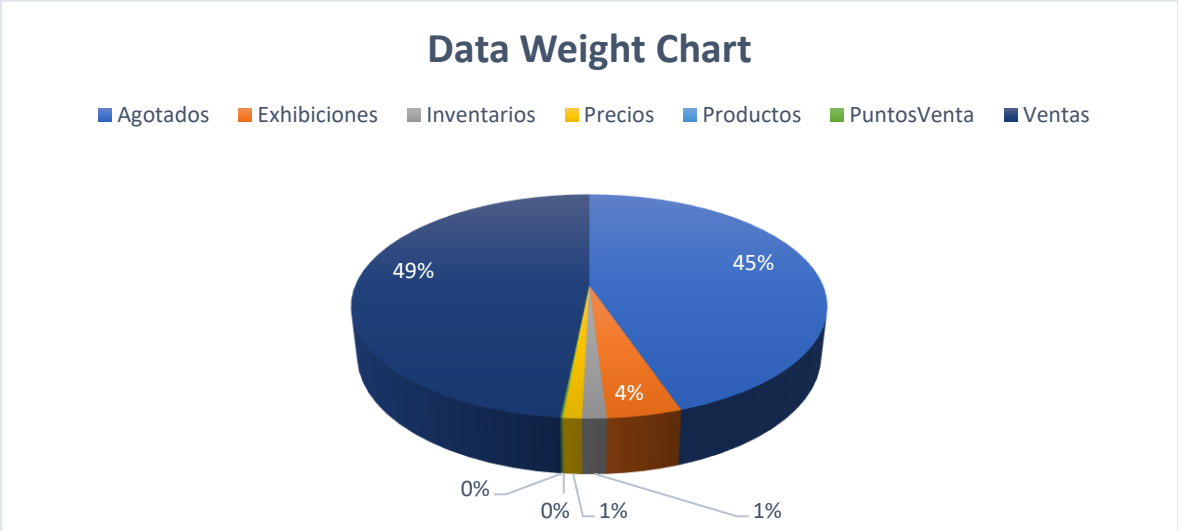


*Figure 2 Data size pie chart*

## 1.1 PRODUCTOS DATA (pro)

The table “productos” as its name indicates contain the information related to all the products that are sold according to the provide by the customer.

The table “productos” initially have 7 columns in order to simplify and standardize the data we proceed to update the column names and assign a more appropriate data type in order to facilitate the future analysis, also columns Unnamed and Linea are deleted the first one is an index which is not need since the IdProducto is the primary key and unique index for each row and Linea is deleted since it contains duplicated information that is already contained in “categoria”:

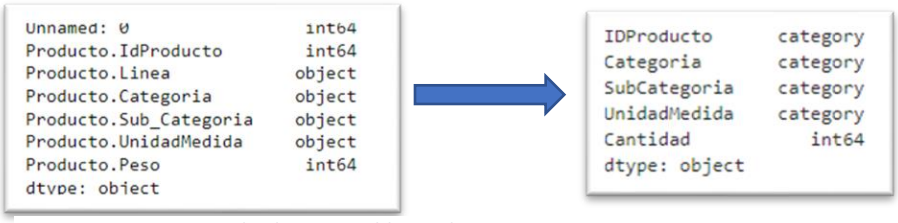


*Figure 3. Standardization Table Productos*

Once the standardization process is completed, we proceed to create the data frame and display the 3 first rows shown on the image below:

| | IDProducto | Linea | Categoria | SubCategoria | UnidadMedida | Peso |
|---|---|---|---|---|---|---|
| **0** | 122902 | Quitagrasa | Quitagrasa | Quitagrasa Pistola | ML | 500 |
| **1** | 122904 | Quitagrasa | Quitagrasa | Quitagrasa Repuesto | ML | 2000 |
| **2** | 122905 | Quitagrasa | Quitagrasa | Multiusos Repuesto | ML | 500 |

*Figure 4 Standardize table "Productos" (pro)*

Metrics for each column:

| Columna | Distintos | Descripcion |
|---|---|---|
| IDProducto | 233 | identificador del producto |
| Linea | 15 | linea a la que pertenece |
| Categoria | 15 | Categoria a la que pertenece |
| SubCategoria | 76 | Sub categoria a la que pertenece |
| UnidadMedida | 4 | unidad de medida del producto |
| Peso | 56 | medida segun cada unidad |

We see that the `Line` and `Category` columns have 15 different values and they seem the same, evaluating that they are equal, we proceed to discard the `Line` column. Additionally we rename the `Peso` column to `Cantidad`

| | IDProducto | Categoria | SubCategoria | UnidadMedida | Cantidad |
|---|---|---|---|---|---|
| **0** | 122902 | Quitagrasa | Quitagrasa Pistola | ML | 500 |
| **1** | 122904 | Quitagrasa | Quitagrasa Repuesto | ML | 2000 |
| **2** | 122905 | Quitagrasa | Multiusos Repuesto | ML | 500 |

Next step, some categorical variables are graphed to visualize the information more easily.

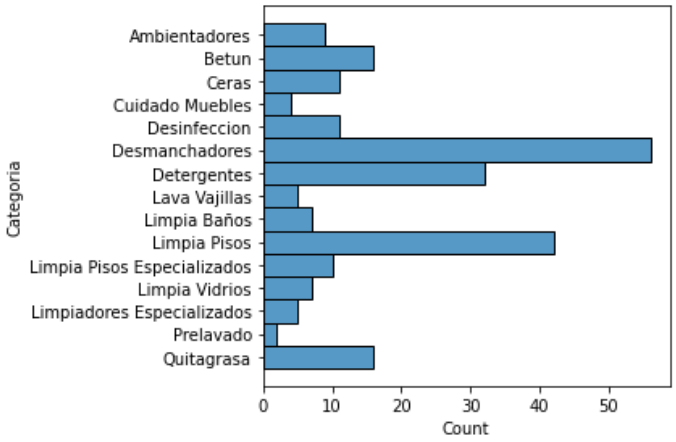


*Figure 5 Frequency table grouped by Category*

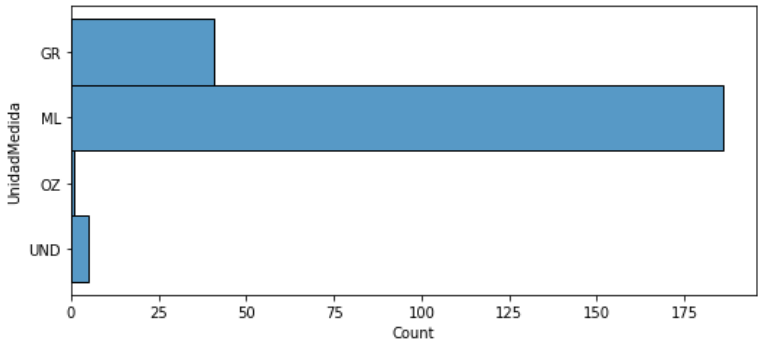


*Figure 6 Frequency by Measure Unit*

As an initial conclusion of this dataframe, it is found that there are 233 different products, grouped into 15 categories and 76 subcategories, with two additional columns that serve as descriptors of the product in question plus the quantity.

### a) Review of N/A values on table productos

This is to confirm that there are not N/A values that could potentially distorsionate the data contain on the table and the future analysis:

```
1 pro.isna().any()
```

```
IDProducto      False
Linea           False
Categoria       False
SubCategoria    False
UnidadMedida    False
Peso            False
dtype: bool
```

### b) Identification of unique values for IdProducto, Linea, Categoria y Subcategoria

To be able to perform join and merge operations with this tables to others we consider important to find the unique values for the index in this table, which is the IdProducto, Linea, Categoria y Subcategoria

```
1 print(pro['IDProducto'].unique())
```

```
['122902', '122904', '122905', '122907', '122908', ..., '122903', '122906', '129214', '122954', '122955']
Length: 233
Categories (233, object): ['122902', '122904', '122905', '122907', ..., '122906', '129214', '122954', '122955']
```

```
1 print(pro['Linea'].unique())
```

```
['Quitagrasa' 'Limpia Vidrios' 'Limpia Baños' 'Detergentes'
 'Lava Vajillas' 'Limpia Pisos Especializados' 'Cuidado Muebles'
 'Limpiadores Especializados' 'Desmanchadores' 'Limpia Pisos' 'Ceras'
 'Betun' 'Ambientadores' 'Desinfeccion' 'Prelavado']
```

```
1 print(pro['Categoria'].unique())
```

```
['Quitagrasa' 'Limpia Vidrios' 'Limpia Baños' 'Detergentes'
 'Lava Vajillas' 'Limpia Pisos Especializados' 'Cuidado Muebles'
 'Limpiadores Especializados' 'Desmanchadores' 'Limpia Pisos' 'Ceras'
 'Betun' 'Ambientadores' 'Desinfeccion' 'Prelavado']
```

```
1 print(pro['SubCategoria'].unique())
```

```
['Quitagrasa Pistola' 'Quitagrasa Repuesto' 'Multiusos Repuesto'
 'Multiusos Pistola' 'Vidrios Repuesto' 'Vidrios Pistola' 'Liquido'
 'Pastillas Blanco' 'Pastillas Azules' 'Ropa Oscura 1800ml'
 'Todos Los Dias 1800ml' 'Todos Los Dias 3785ml' 'Ropa Oscura 3785ml'
 'Ropa Bebe 900ml' 'Ropa Oscura 500ml' 'Ropa Oscura 5000ml'
 'Todos Los Dias 5000ml' 'Ropa Oscura 900ml' 'Todos Los Dias 900ml'
 'Todos Los Dias 300ml' 'Ropa Bebe 2000ml' 'Tabletas' 'Abrillantador'
 'Todos Los Dias 3785ml + 900 Ml' 'Ropa Oscura 3785ml + 900 Ml'
 'Todos Los Dias 1000ml' 'Todos Los Dias 2000ml'
 'Todos Los Dias 3785ml X2' 'Todos Los Dias 1000ml + 2000ml'
 'Ropa Oscura 1000ml' 'Ropa Oscura 2000ml' 'Ropa Oscura 3785ml X2'
 'Ropa Bebe 1000ml' 'Polvo' 'Todos Los Dias 5400ml'
 'Todos Los Dias 3600ml' 'Ropa Bebe 3785ml' 'Pisos' 'Cubrerasguños'
 'Lustramuebles' 'Bronce Y Otros Pulidores De Metales'
 'Limpiadores De Hornos' 'Pulidores De Metales En Plata' 'Liquido Blanco'
 'Liquido Rosa' 'Polvo Blanco' 'Polvo Rosa' 'Alegria' 'Energia'
 'Tranquilidad' 'Canela Manzana' 'Vainilla' 'Lavanda' 'Prelavado Pistola'
 'Prelavado Pistola + Repuesto' 'Liquido Blanco Y Rosa'
 'Prelavado Repuesto' 'Perfumados' 'Especializados' 'Brillo Instantaneo'
 'Maximo Brillo' 'Sanpic Mascotas' 'Pasta 96gr Negro' 'Cremoso 100ml'
 'Pasta 65gr Cherry' 'Pasta 30gr Cherry' 'Liquidos Cherry'
 'Pasta 12gr Cherry' 'Liquidos Griffin' 'Ambientadores Electricos'
 'Ambientadores Fm' 'Repuesto X 2' 'Aerosoles' 'Pañitos' 'Lysol Cocina']
```

## 1.2 AGOTADOS DATA (ago)

The table “Agotados” provide us the information regarding the out of stock products on this table we can see that there is a foreign key coming from the previous table productos which allows us to identify information such as the out of stock product id, the cause of the out of stock, date, among others.

Next, the summary of the table after a first delimiting of data types by columns:

| Column | Unique | Description |
|---|---|---|
| IDRegistro | 1364262 | sample registration |
| Agotado | 2 | product is available |
| Presente | 2 | Product present at the point of sale |
| NOManejo | 2 | Product that is not handled or is not coded at the Point of Sale |
| Causal | 1 | Main out of stock cause |
| SubCausal | 8 | Secondary out of stock cause |
| IDProducto | 229 | Product identifier |
| IDPdv | 704 | Point of Sale Identifier |
| IDUser | 97 | User identifier |
| Fecha | 194 | sample date |

First five rows:

| | IDRegistro | Agotado | Presente | NOManejo | Causal | SubCausal | IDProducto | IDPdv | IDUser | Fecha |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 313610627 | False | True | False | NaN | NaN | 122894 | 1597217 | 123008 | 2021-09-27 |
| **1** | 313610624 | False | True | False | NaN | NaN | 122921 | 1597217 | 123008 | 2021-09-27 |
| **2** | 313610625 | False | True | False | NaN | NaN | 123072 | 1597217 | 123008 | 2021-09-27 |
| **3** | 313610626 | False | True | False | NaN | NaN | 123033 | 1597217 | 123008 | 2021-09-27 |
| **4** | 313610628 | False | True | False | NaN | NaN | 122959 | 1597217 | 123008 | 2021-09-27 |

In the following table we can identify the data types of the dataset:

```
IDRegistro          category
Agotado                 bool
Presente                bool
NOManejo                bool
Causal                object
SubCausal             object
IDProducto          category
IDPdv               category
IDUser              category
Fecha         datetime64[ns]
```

Statistical summary:

| | count | unique | top | freq | first | last |
|---|---|---|---|---|---|---|
| **IDRegistro** | 1364563 | 1364262 | 3.41e+08 | 82 | NaT | NaT |
| **Agotado** | 1364563 | 2 | False | 1299160 | NaT | NaT |
| **Presente** | 1364563 | 2 | True | 1257143 | NaT | NaT |
| **NOManejo** | 1364563 | 2 | False | 1322546 | NaT | NaT |
| **Causal** | 65505 | 1 | Agotado | 65505 | NaT | NaT |
| **SubCausal** | 65505 | 8 | Pedido Insuficiente | 40009 | NaT | NaT |
| **IDProducto** | 1364563 | 229 | 123082 | 14069 | NaT | NaT |
| **IDPdv** | 1364563 | 704 | 1596708 | 12402 | NaT | NaT |
| **IDUser** | 1364563 | 97 | 123563 | 66951 | NaT | NaT |
| **Fecha** | 1364563 | 194 | 2022-03-28 00:00:00 | 12829 | 2021-09-27 | 2022-05-21 |

Also, we check the percentage of null data in the dataset:

```
IDRegistro     0.0
Agotado        0.0
Presente       0.0
NOManejo       0.0
Causal        95.2
SubCausal     95.2
IDProducto     0.0
IDPdv          0.0
IDUser         0.0
Fecha          0.0
```

From the above result, we identify that exist a huge number of non-values in the Columns Causal and SubCausal. In the following step, we will identify deeper about the relation of this result with the Agotados column.

We can look for a distribution in the columns Agotado and Fecha with a bar graph, we identify that the quantity of out-of-stock products have almost the same frequency of the Causal and SubCausal columns, so they are related and important for the Advance EDA.

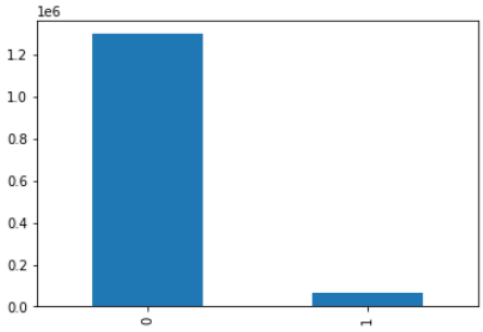


*Figure 7 False and True Frequency for Agotados*

We see then the huge difference between these two values in the same column (False 1’299.160 and True 65.403).

The following is a graph of the count measures for each month. In this moment, this does not tell us too much about the behavior of the data, we should identify the products out-of-stock, filter by them, by location (City, Depto, Region and City) and more features looking for relevant information in the advance EDA.

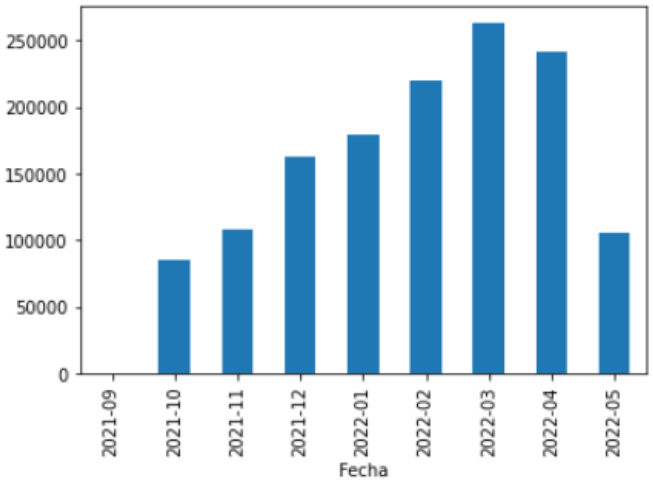


*Figure 8 Frequency Table Monthly Quantity*

## 1.3 EXHIBICIONES DATA (exh)

The table “Exhibiciones” contain the information related to the products that are currently being exhibited on the different sales points on this table we can see that we have foreign keys to identify important information such as de Point of Sale identification and the product category that will later help us to identify the relation between the out of stock and the exhibitions.

This table have the following columns and description:

| Column | Unique | Description |
|---|---|---|
| IDPdv | 693 | Point of Sale Identifier |
| Fecha | 157 | Sample date |
| IDRegistro | 13144 | Sample Registration |
| Categoria | 12 | Product category |
| Implementacion | 84389 | The implementation is done |
| Causal | 4 | Main reason for no implementation |
| POP | 2 | There is publicity material |
| BloqueMarca | 2 | There is a block for the product |
| OpeLogis | 3 | Logistic operator indentifier |

First five rows of the Dataframe

| | IDPdv | Fecha | IDRegistro | Categoria | Implementacion | Causal | POP | BloqueMarca | OpeLogis |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1596927 | 2021-11-30 | 1596927-127115-2021/11/30 | DESINFECCION | True | NaN | False | False | 2 |
| **1** | 1596627 | 2021-11-26 | 1596627-121221-2021/11/26 | BETUN | True | NaN | False | True | 2 |
| **2** | 1596627 | 2021-11-26 | 1596627-121221-2021/11/26 | BETUN | True | NaN | False | True | 2 |
| **3** | 1597233 | 2021-11-29 | 1597233-123565-2021/11/29 | DESMANCHADORES | True | NaN | True | True | 2 |
| **4** | 1596627 | 2021-11-30 | 1596627-123565-2021/11/30 | DESINFECCION | False | BAJO INVENTARIO | True | False | 1 |

Datatypes

```
IDPdv                category
Fecha          datetime64[ns]
IDRegistro           category
Categoria              object
Implementacion         object
Causal                 object
POP                    object
BloqueMarca            object
OpeLogis             category
```

Statistical summary

| | count | unique | top | freq | first | last |
|---|---|---|---|---|---|---|
| **IDPdv** | 87357 | 693 | 1597201 | 610 | NaT | NaT |
| **Fecha** | 87357 | 157 | 2022-04-04 00:00:00 | 940 | 2021-11-13 | 2022-05-23 |
| **IDRegistro** | 87357 | 13144 | 1597312-141223-2022/04/21 | 133 | NaT | NaT |
| **Categoria** | 87357 | 12 | DESMANCHADORES | 13606 | NaT | NaT |
| **Implementacion** | 84389 | 2 | True | 78167 | NaT | NaT |
| **Causal** | 6604 | 4 | BAJO INVENTARIO | 2779 | NaT | NaT |
| **POP** | 84080 | 2 | False | 66697 | NaT | NaT |
| **BloqueMarca** | 84080 | 2 | True | 76290 | NaT | NaT |
| **OpeLogis** | 58326 | 3 | 2 | 46145 | NaT | NaT |

This dataset has information about 693 IDPdv compared to the total 1702 from the dataset IDPdv.

```
IDPdv             0.0
Fecha             0.0
IDRegistro        0.0
Categoria         0.0
Implementacion    3.4
Causal           92.4
POP               3.8
BloqueMarca       3.8
OpeLogis         33.2
```

Here we see some null data in the columns, but in future steps it will be clearly if those columns add information for the objective of project or not.

We can search for distribution in the columns. With a value_counts(), it is easier to see the range of data from each IDPdv, this will be crucial for the data analysis due to the differences in the values from each IDPdv.

| | |
|---|---|
| count | 693.000000 |
| mean | 126.056277 |
| std | 105.282546 |
| min | 1.000000 |
| 25% | 47.000000 |
| 50% | 104.000000 |
| 75% | 171.000000 |
| max | 610.000000 |

For the Dates, we also see a different behavior for each month, this also generates a behavior in the data that we must describe for a better understanding.

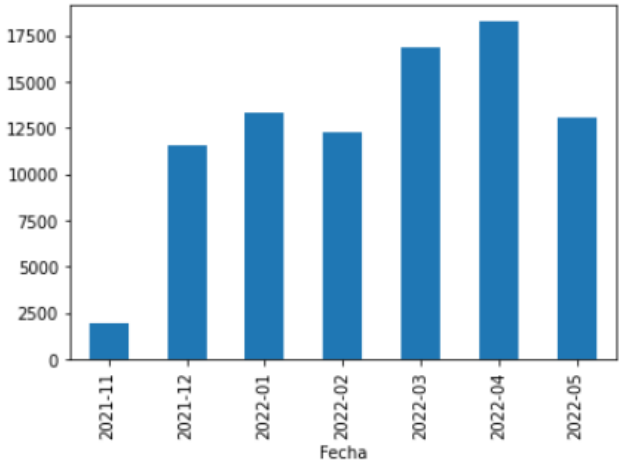


*Figure 9 Exhibitions Frequency Table per Month*

This column will be compared to the same in the dataset of Products, but in this moment it is not possible to discard this one from the dataset.

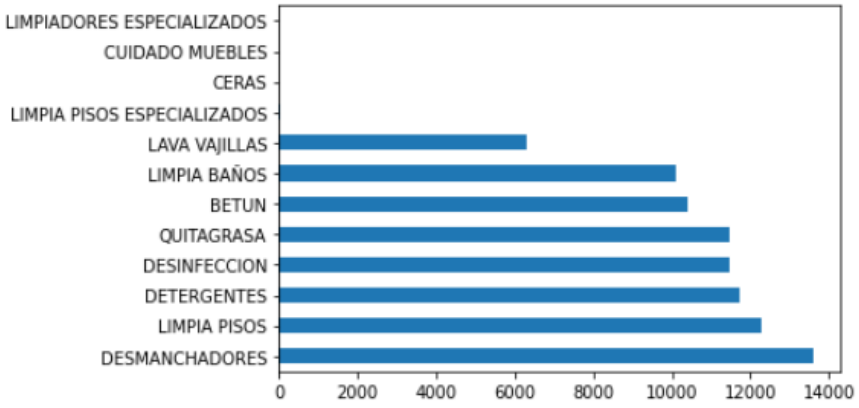


*Figure 10 Frequency Table of Categoria*

In conclusion, this table give us information about the exhibitions on each point of sale, but we will have to check if this data can create useful information for the project problem in future steps.

## 1.4 INVENTARIO DATA (inv)

This table have the following columns and description:

| Column | Unique | Description |
|---|---|---|
| Fecha | 148 | Sample date |
| IDPdv | 1462 | Point of sale identifier |
| IDProducto | 199 | Product Identifier |
| StockUnidades | -- | Average daily products |
| StockTrans | -- | Average of products until supply |

We can see that StockUnidades and StockTrans are columns with incorrect values, looking closely to these, we found that there are 735 negative values mixed with the data:

We must find the reason for this data in future steps or decide to delete this data depending on what additional information is possible to analyse.

First five rows of the dataset:

| | Fecha | IDPdv | IDProducto | StockUnidades | StockTrans |
|---|---|---|---|---|---|
| **0** | 2021-10-03 | 1596624 | 122907 | 0 | 0 |
| **1** | 2021-10-03 | 1596633 | 123048 | 0 | 0 |
| **2** | 2021-10-03 | 1596629 | 122938 | 0 | 0 |
| **3** | 2021-10-03 | 1614610 | 122894 | 0 | 0 |
| **4** | 2021-10-03 | 1614610 | 122893 | 0 | 0 |

Datatypes

```
Fecha            datetime64[ns]
IDPdv                  category
IDProducto             category
StockUnidades             int64
StockTrans                int64
```

Statistical summary

| | count | unique | top | freq | first | last |
|---|---|---|---|---|---|---|
| **Fecha** | 62166 | 148 | 2022-05-19 00:00:00 | 46663 | 2021-04-02 | 2022-05-19 |
| **IDPdv** | 62166 | 1462 | 7701008009196 | 129 | NaT | NaT |
| **IDProducto** | 62166 | 199 | 123042 | 944 | NaT | NaT |
| **StockUnidades** | 62166.0 | NaN | NaN | NaN | NaT | NaT |
| **StockTrans** | 62166.0 | NaN | NaN | NaN | NaT | NaT |

Distributions

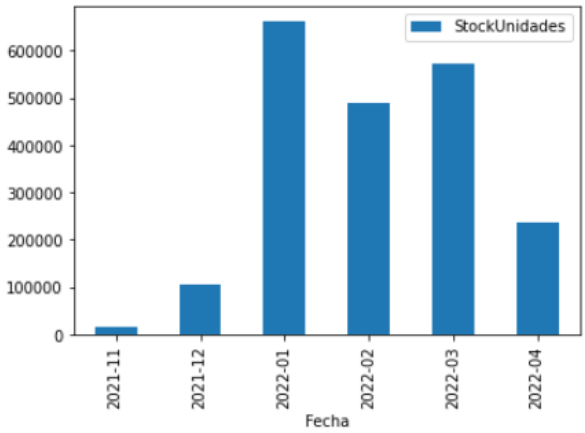


*Figure 11 Total Stock per Months*

The previous bar graph tells us that the behavior of the data is not stable or maybe it is affected by an event that made the collected data lesser to other months.

Since the inventory table is one of the most important because the project will be focusing on out of stock and how to predict them, we perform several process to clean the data contain on this table to later have an easiest way to analyze the info

### a) Review of N/A values

This is to confirm that there are not N/A values that could potentially distorsionate the data contain on the table and the future analysis:

```
In [15]: 1 inv.isna().any()

Out[15]: Fecha           False
         IDPdv           False
         IDProducto      False
         StockUnidades   False
         StockTrans      False
         dtype: bool
```

### b) Negative Values on Quantity Column Table Inventarios

We could identify that there were negative values on the column quantity in inventories we confirm about this with the client since it doesn't make to have a variable such as quantity with negative values and they confirm that this could be either due to mistaken information or inventory shortages, in order to avoid any misleading on the analysis due to this information we deleted from the table after this was identify

```
1 inv.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62166 entries, 0 to 62165
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Fecha          62166 non-null  datetime64[ns]
 1   IDPdv          62166 non-null  category
 2   IDProducto     62166 non-null  category
 3   StockUnidades  62166 non-null  int64
 4   StockTrans     62166 non-null  int64
dtypes: category(2), datetime64[ns](1), int64(2)
memory usage: 1.7 MB
```

```
1 inv[inv.StockUnidades < 0].shape[0]

735
```

```
inv.drop(inv[inv['StockUnidades'] < 0].index, inplace = True)
print(inv)
```

```
           Fecha          IDPdv IDProducto  StockUnidades  StockTrans
0     2021-10-03        1596624     122907              0           0
1     2021-10-03        1596633     123048              0           0
2     2021-10-03        1596629     122938              0           0
3     2021-10-03        1614610     122894              0           0
4     2021-10-03        1614610     122893              0           0
...          ...            ...        ...            ...         ...
62161 2022-05-19  7701001040578     122941              6           0
62162 2022-05-19  7701001040578     122940             15           0
62163 2022-05-19  7701001040578     122906             15           0
62164 2022-05-19  7701001040578     122907             30           0
62165 2022-05-19  7701001040271     122964             36           0

[61431 rows x 5 columns]
```

```
# Se realiza verificación de que las 735 filas con datos negativos hayan sido eliminadas
inv.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 61431 entries, 0 to 62165
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Fecha          61431 non-null  datetime64[ns]
 1   IDPdv          61431 non-null  category
 2   IDProducto     61431 non-null  category
 3   StockUnidades  61431 non-null  int64
 4   StockTrans     61431 non-null  int64
dtypes: category(2), datetime64[ns](1), int64(2)
memory usage: 2.2 MB
```

### c) Association “Categoria” from table “productos” to table “Inventarios”

The table inventarios does not contain information that allow us to group the data into similar categories for that reason and after identifying the different categories the products have on the table “productos” column “categoria” we associate it into the table inventarios creating a new column and using a merge operation

```
pro_categoria = pro[['IDProducto','Linea']].copy()
pro_categoria.head()
```

| | IDProducto | Linea |
|---|---|---|
| 0 | 122902 | Quitagrasa |
| 1 | 122904 | Quitagrasa |
| 2 | 122905 | Quitagrasa |
| 3 | 122907 | Quitagrasa |
| 4 | 122908 | Quitagrasa |

```
inv = pd.merge(
                inv,
                pro_categoria,
                on = 'IDProducto',
                how = 'inner')
inv.head()
```

| | Fecha | IDPdv | IDProducto | StockUnidades | StockTrans | Mes | Depto | Linea |
|---|---|---|---|---|---|---|---|---|
| 0 | 2021-10-03 | 1596624 | 122907 | 0 | 0 | 2021-10 | NARIÑO | Quitagrasa |
| 1 | 2021-08-02 | 1614355 | 122907 | 0 | 0 | 2021-08 | VALLE DEL CAUCA | Quitagrasa |
| 2 | 2022-05-19 | 1599404 | 122907 | 3 | 0 | 2022-05 | LA GUAJIRA | Quitagrasa |
| 3 | 2022-05-19 | 1596792 | 122907 | 60 | 0 | 2022-05 | SANTANDER | Quitagrasa |
| 4 | 2022-05-19 | 1596810 | 122907 | 15 | 0 | 2022-05 | NORTE DE SANTANDER | Quitagrasa |

### c) Association “Departamentos” from table “pdv” to table “Inventarios”

In order to facilitate a geographical analysis, we take the “departamento” column coming from the “puntos de venta” table and associate it so we can analyze the behavior of the total stock by departments and category

```
print(pdv['Depto'].unique())
```

```
['ANTIOQUIA' 'BOGOTA D.C' 'VALLE DEL CAUCA' 'HUILA' 'ATLÁNTICO'
 'LA GUAJIRA' 'CASANARE' 'SANTANDER' 'NARIÑO' 'NORTE DE SANTANDER'
 'RISARALDA' 'CUNDINAMARCA' 'BOYACA' 'CORDOBA' 'CESAR' 'BOLIVAR'
 'MAGDALENA' 'CAUCA' 'TOLIMA' 'CALDAS' 'META' 'QUINDÍO' 'SUCRE' 'CHOCO'
 'ARAUCA' 'CAQUETA']
```

```
pdv.shape
```

```
(1702, 15)
```

```
pdv.head()
```

| | IDPdv | Ciudad | Depto | Cadena | Canal | SubCanal | Tipologia | Regional | Formato | DiaEntregaPedidoNombre | DiaEntregaPedido | DiaHoy | Dia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1596878 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | |
| 1 | 1596877 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | |
| 2 | 1596876 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | |
| 3 | 1596875 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | |
| 4 | 1614373 | Bogotá | BOGOTA D.C | Cencosud | Mt | Ka | CONVENIENCE | Cundi-Boyaca | Metro Express | NaN | 0 | 1 | |

```
pdv_region = pdv[['IDPdv','Depto']].copy()
pdv_region.head()
```

| | IDPdv | Depto |
|---|---|---|
| 0 | 1596878 | ANTIOQUIA |
| 1 | 1596877 | ANTIOQUIA |
| 2 | 1596876 | ANTIOQUIA |
| 3 | 1596875 | ANTIOQUIA |

```
inv = pd.merge(
                inv,
                pdv_region,
                on = 'IDPdv',
                how = 'inner')
inv.head()
```

| | Fecha | IDPdv | IDProducto | StockUnidades | StockTrans | Mes | Depto |
|---|---|---|---|---|---|---|---|
| 0 | 2021-10-03 | 1596624 | 122907 | 0 | 0 | 2021-10 | NARIÑO |
| 1 | 2021-10-10 | 1596624 | 125046 | 0 | 0 | 2021-10 | NARIÑO |
| 2 | 2021-10-24 | 1596624 | 125050 | 0 | 0 | 2021-10 | NARIÑO |
| 3 | 2021-10-24 | 1596624 | 125047 | 0 | 0 | 2021-10 | NARIÑO |
| 4 | 2021-10-31 | 1596624 | 125002 | 0 | 0 | 2021-10 | NARIÑO |

### d) Histogram of Average Stock total country (Colombia)

After having the products information complemented with the category so we can group the data and also after including the department we proceed to prepare an histogram by categories with the total stock per all the dates provided on the inventarios table

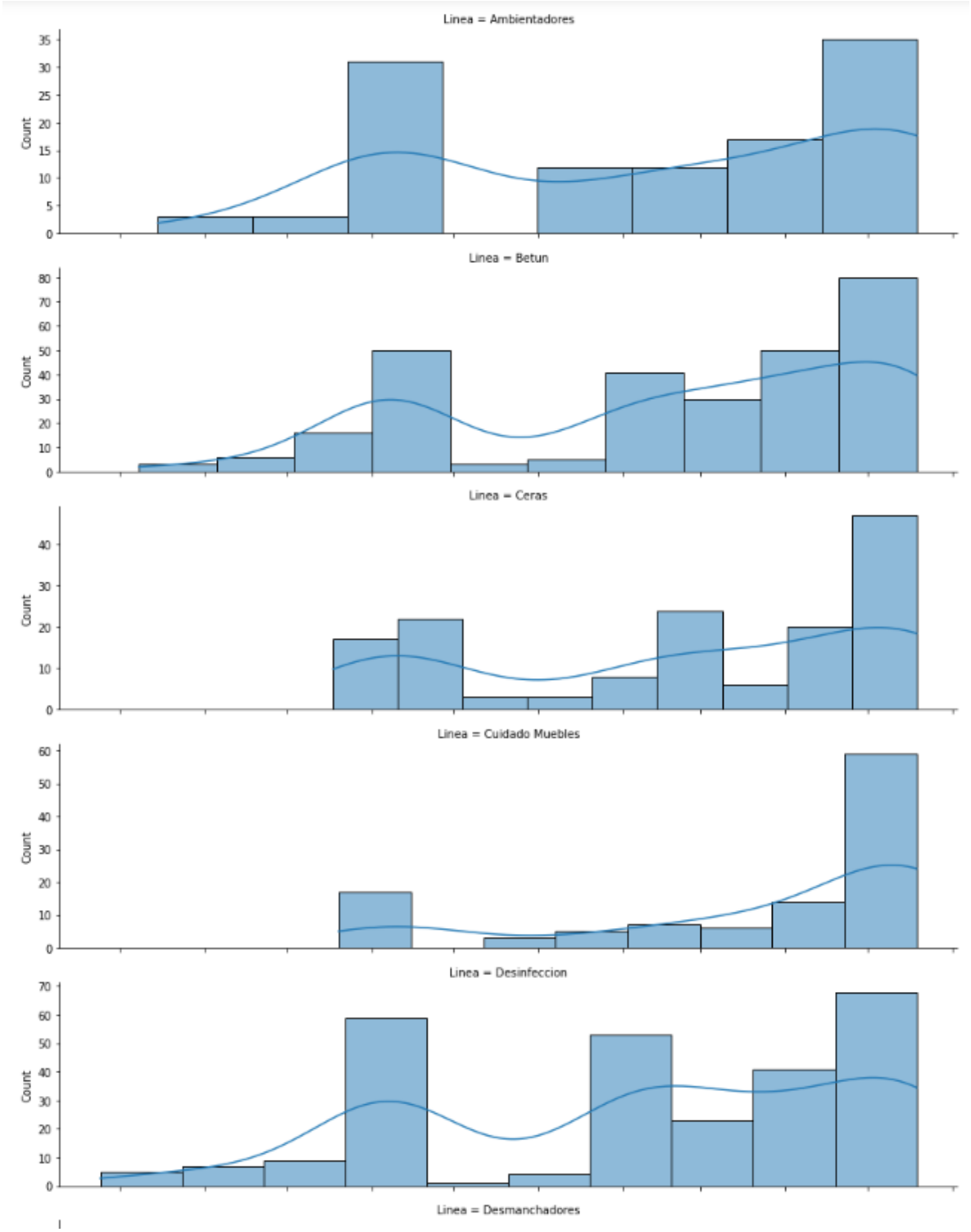


*Figure 12 Histogram Stock by Categorie Per Dates Nation Wide I*

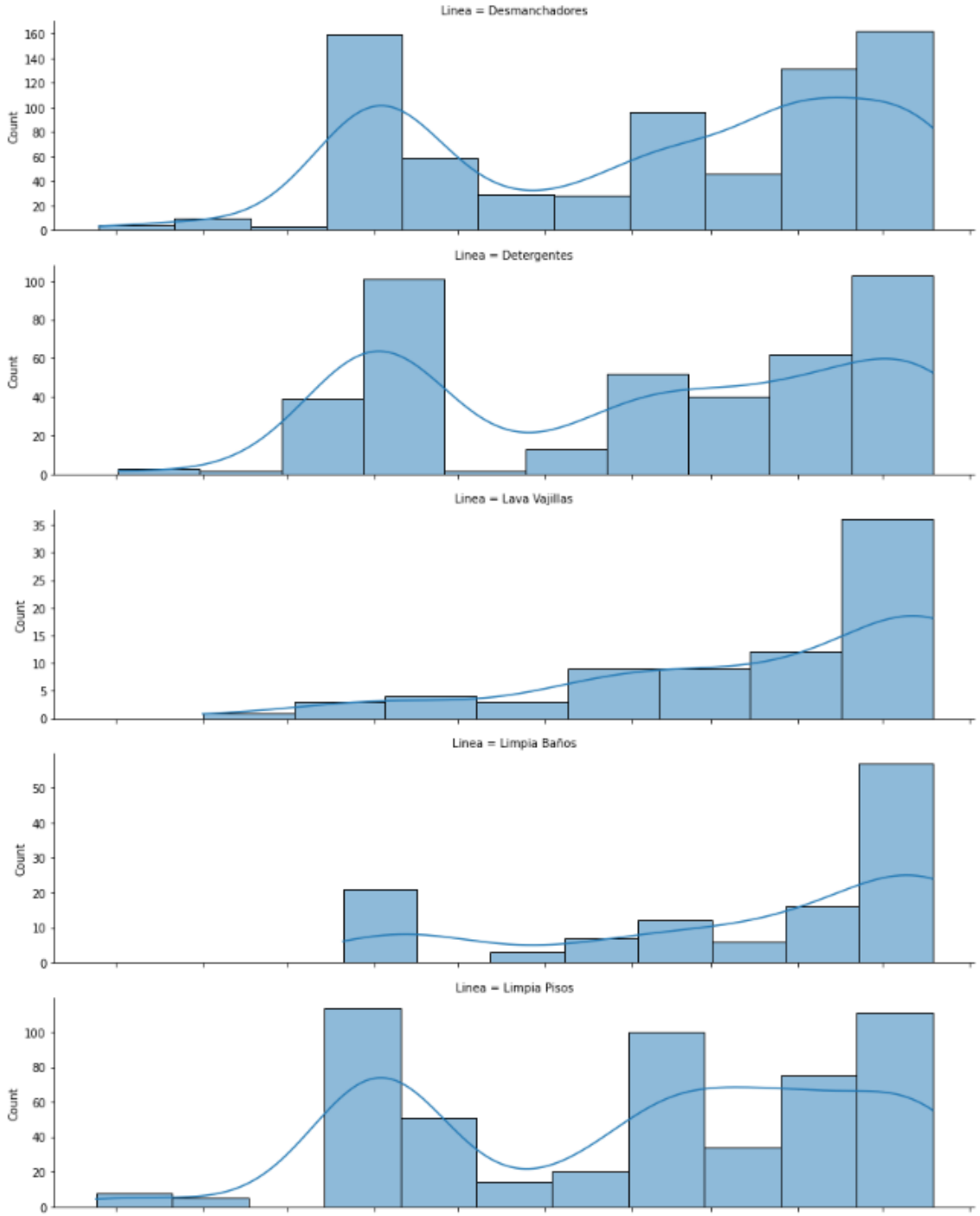


*Figure 13Histogram Stock by Categorie Per Dates Nation Wide II*

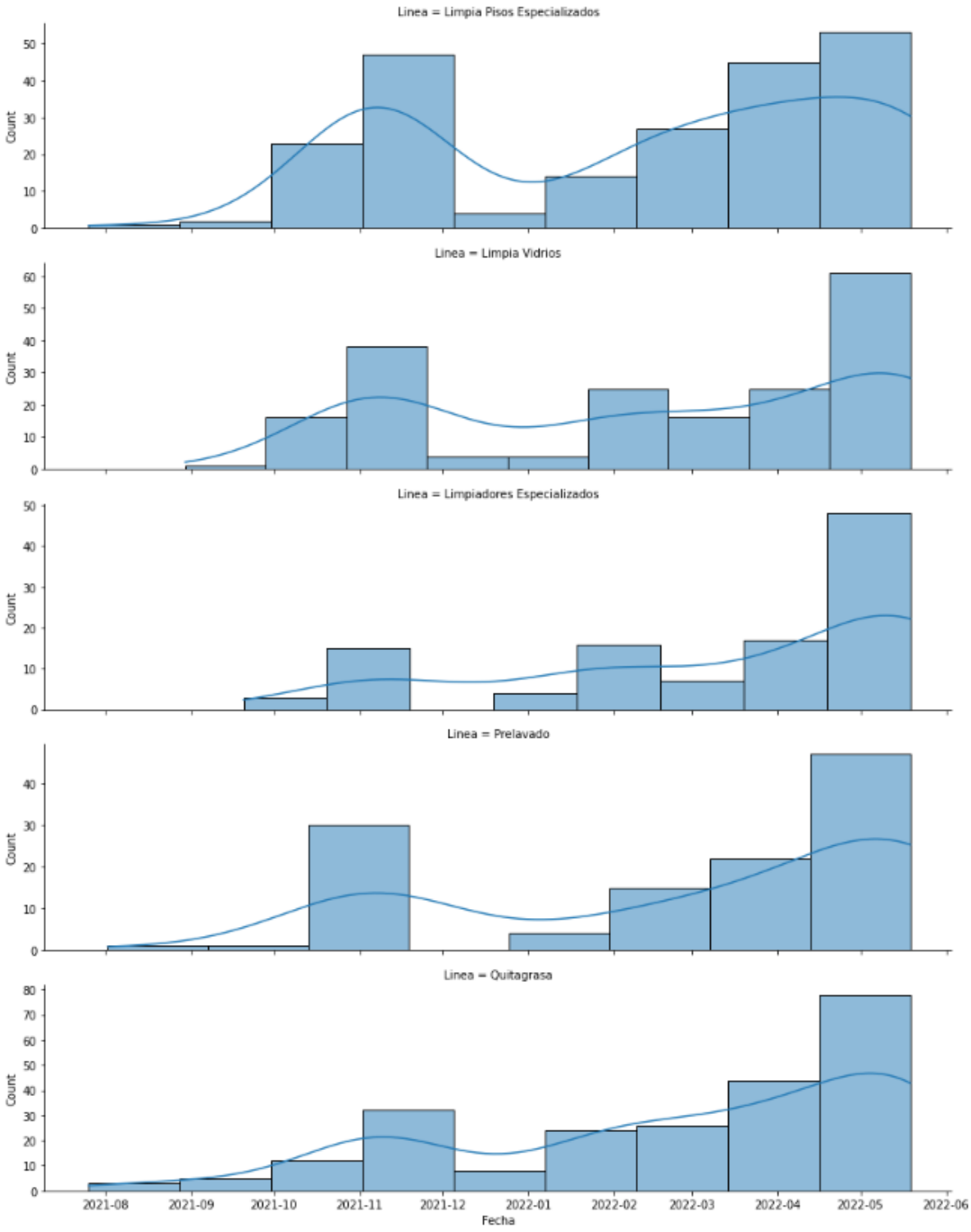


*Figure 14Histogram Stock by Categorie Per Dates Nation Wide III*

## 1.5 PRECIOS DATA (pre)

These are the columns in the dataset.

| Column | Unique | Description |
|---|---|---|
| IDRuta | 680 | Route identifier |
| IDPdv | 479 | Sales point identifier |
| IDProducto | 135 | Product Identifier |
| Valor | 1861 | Product's value |
| Fecha | 184 | Sample date |
| PreSugerido | 153 | Suggested price |

First five rows.

| | IDRuta | IDPdv | IDProducto | Valor | Fecha | PreSugerido |
|---|---|---|---|---|---|---|
| **0** | 1596677123545 | 1596677 | 122965 | 9990 | 2021-10-01 | 11400 |
| **1** | 1596677123545 | 1596677 | 123072 | 6190 | 2021-10-01 | 7950 |
| **2** | 1596677123545 | 1596677 | 122945 | 2990 | 2021-10-01 | 37300 |
| **3** | 1596677123545 | 1596677 | 123033 | 13990 | 2021-10-01 | 16750 |
| **4** | 1596677123545 | 1596677 | 122959 | 37390 | 2021-10-01 | 47800 |

Data types.

```
IDRuta              category
IDPdv               category
IDProducto          category
Valor                  int64
Fecha          datetime64[ns]
PreSugerido            int64
```

Statistical summary.

| | count | unique | top | freq | first | last | mean |
|---|---|---|---|---|---|---|---|
| **IDRuta** | 32287 | 680 | 1612656123584 | 698 | NaT | NaT | NaN |
| **IDPdv** | 32287 | 479 | 1612656 | 970 | NaT | NaT | NaN |
| **IDProducto** | 32287 | 135 | 122907 | 559 | NaT | NaT | NaN |
| **Valor** | 32287.0 | NaN | NaN | NaN | NaT | NaT | 53979.713073 |
| **Fecha** | 32287 | 184 | 2021-10-23 00:00:00 | 1255 | 2021-10-01 | 2022-05-23 | NaN |
| **PreSugerido** | 32287.0 | NaN | NaN | NaN | NaT | NaT | 13610.665903 |

After looking the data, we found outliners in the column Valor, so looking deeper, we see that the 99% quantile is 43.863 pesos, we filter for values greater than 50.000:

| | IDRuta | IDPdv | IDProducto | Valor | Fecha | PreSugerido |
|---|---|---|---|---|---|---|
| **343** | 1597351123585 | 1597351 | 123057 | 116090 | 2021-10-02 | 16200 |
| **2168** | 1597172123585 | 1597172 | 123081 | 318000 | 2021-10-11 | 26000 |
| **3717** | 1596642123585 | 1596642 | 123049 | 65900 | 2021-10-22 | 17100 |
| **4859** | 1597004123556 | 1597004 | 123054 | 64500 | 2021-10-23 | 5500 |
| **4871** | 1597004123556 | 1597004 | 123055 | 64500 | 2021-10-23 | 5500 |
| **...** | ... | ... | ... | ... | ... | ... |
| **31361** | 1596668123579 | 1596668 | 122959 | 51990 | 2022-05-05 | 47800 |
| **31362** | 1596685147042 | 1596685 | 122959 | 51990 | 2022-05-19 | 47800 |
| **31363** | 1596987123579 | 1596987 | 122959 | 57300 | 2022-05-10 | 47800 |
| **31365** | 1597166142752 | 1597166 | 122959 | 50400 | 2022-05-04 | 47800 |
| **31697** | 1596731135753 | 1596731 | 122901 | 115090 | 2022-05-09 | 9550 |

121 rows × 6 columns

It is decided to eliminate the 121 columns that are outside the 99% quantile and the following is the result of the distribution.

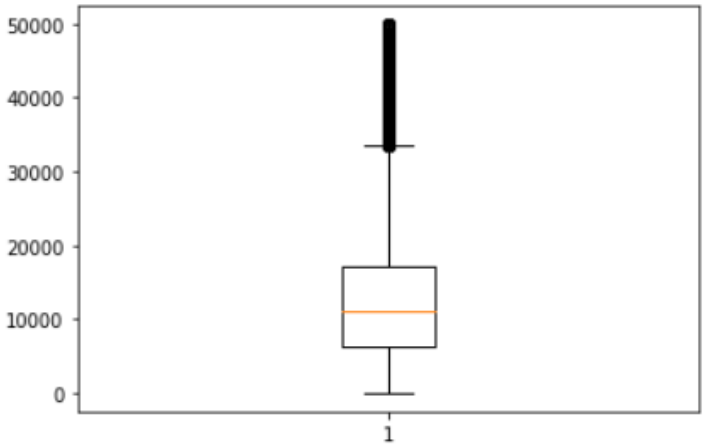


*Figure 15 Boxplot of the Valor Column*

The Precios dataframe has 135 products with a price (98 less than in the Productos table), in 479 different points of sale, this information was taken 184 times, in a period of 234 days from October 1, 2021 to May 23, 2022.

121 rows with an abnormally high price range were found, these values were eliminated considering that they correspond to less than 1% of the dataset. It remains for further analysis to review the difference between the Value column and the Suggested Price.

## 1.6 PUNTOS DE VENTA DATA (pdv)

These are the columns in the dataset

| Column | Unique | Description |
|---|---|---|
| IDPdv | 1702 | Point of Sale Identifier |
| Ciudad | 160 | City |
| Depto | 26 | Department |
| Cadena | 87 | chain store |
| Canal | 3 | Sales channel |
| SubCanal | 6 | Sub-sales channel |
| Tipologia | 22 | Tipology of the sales point |
| Regional | 8 | Region |
| Formato | 48 | Channel format |
| DiaEntregaPedidoNombre | 11 | Days of product delivery |
| DiaEntregaPedido | 7 | Order delivery day |
| DiaHoy | 1 | Actual day |
| DiaRestante | 7 | Days left for new delivery |
| Lat | -- | Latitude |
| Lon | -- | Longitude |

First five rows.

| | IDPdv | Ciudad | Depto | Cadena | Canal | SubCanal | Tipologia | Regional | Formato | DiaEntregaPedidoNombre | DiaEntregaPedido | DiaHoy | DiaRestante | Lat | Lon |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1596878 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | 6 | 200.0 | 200.0 |
| 1 | 1596877 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | 6 | 200.0 | 200.0 |
| 2 | 1596876 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | 6 | 200.0 | 200.0 |
| 3 | 1596875 | Medellin | ANTIOQUIA | Cooperativa De Consumo | Dtt | Smi | SMI P | Antioquia-Choco | Surtiventas | NaN | 0 | 1 | 6 | 200.0 | 200.0 |
| 4 | 1614373 | Bogotá | BOGOTA D.C | Cencosud | Mt | Ka | CONVENIENCE | Cundi-Boyaca | Metro Express | NaN | 0 | 1 | 6 | NaN | NaN |

Data types.

```
IDPdv                    category
Ciudad                     object
Depto                      object
Cadena                     object
Canal                      object
SubCanal                   object
Tipologia                  object
Regional                   object
Formato                    object
DiaEntregaPedidoNombre     object
DiaEntregaPedido         category
DiaHoy                   category
DiaRestante              category
Lat                       float64
Lon                       float64
```

Statistical summary

| | count | unique | top | freq | mean |
|---|---|---|---|---|---|
| **IDPdv** | 1702 | 1702 | 1596620 | 1 | NaN |
| **Ciudad** | 1702 | 160 | Bogotá | 491 | NaN |
| **Depto** | 1702 | 26 | BOGOTA D.C | 491 | NaN |
| **Cadena** | 1702 | 87 | Grupo Comercial Éxito | 504 | NaN |
| **Canal** | 1702 | 3 | Mt | 1251 | NaN |
| **SubCanal** | 1702 | 6 | Ka | 1113 | NaN |
| **Tipologia** | 1702 | 22 | SUPERP | 385 | NaN |
| **Regional** | 1702 | 8 | Cundi-Boyaca | 607 | NaN |
| **Formato** | 1702 | 48 | Olimpica Sto | 298 | NaN |
| **DiaEntregaPedidoNombre** | 448 | 11 | Miercoles - Lunes | 103 | NaN |
| **DiaEntregaPedido** | 1702 | 7 | 0 | 1254 | NaN |
| **DiaHoy** | 1702 | 1 | 1 | 1702 | NaN |
| **DiaRestante** | 1702 | 7 | 6 | 1254 | NaN |
| **Lat** | 1135.0 | NaN | NaN | NaN | 2700680.044053 |
| **Lon** | 1135.0 | NaN | NaN | NaN | -31091783.244934 |

From the previous information, we can infer that the columns have information without the presence of null values, except for the Lat and Lon columns. There is a lot of information, but this should be filtered in future steps for its applicability in forecast models.

Distributions.

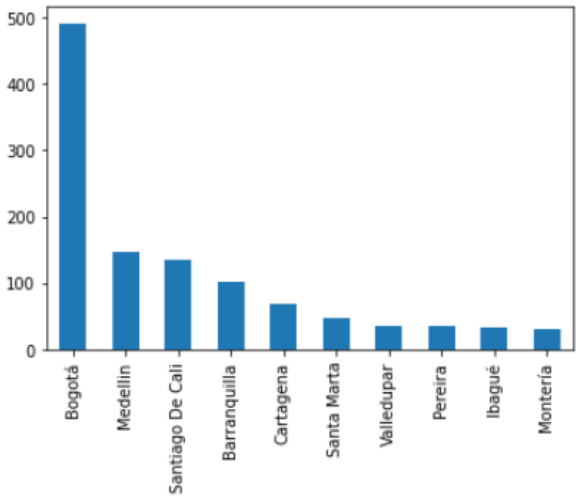


*Figure 16Frequency Table Puntos de Venta per City*

From the previous graph, it is evident that the highest percentage of points of sale are in the city of Bogota, being a reference for the results and conclusions of the dataset study. This also infers that most of the work will be carried out for this city, those that present a considerable volume of data to be able to analyze.

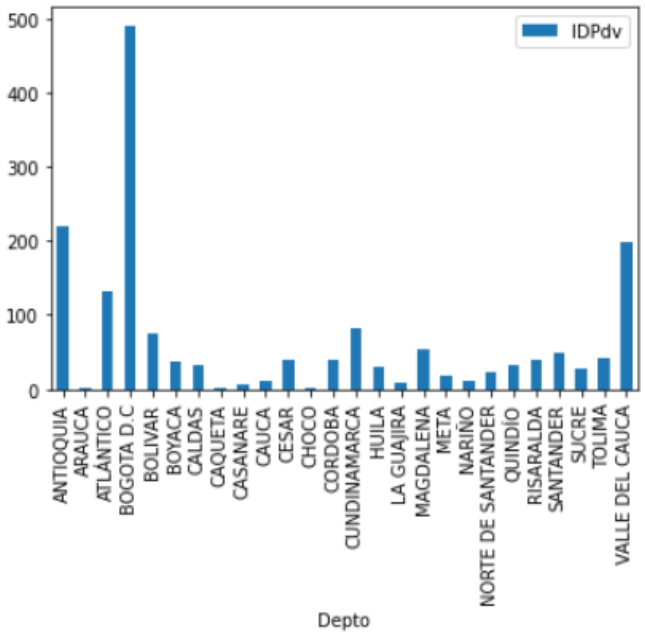


*Figure 17 Frequency Table Puntos de Venta per Department*

This graph confirms the above statement, it should be considered to select only the representative values for further analysis.