Big Data Project: Deliverable 2

Team :

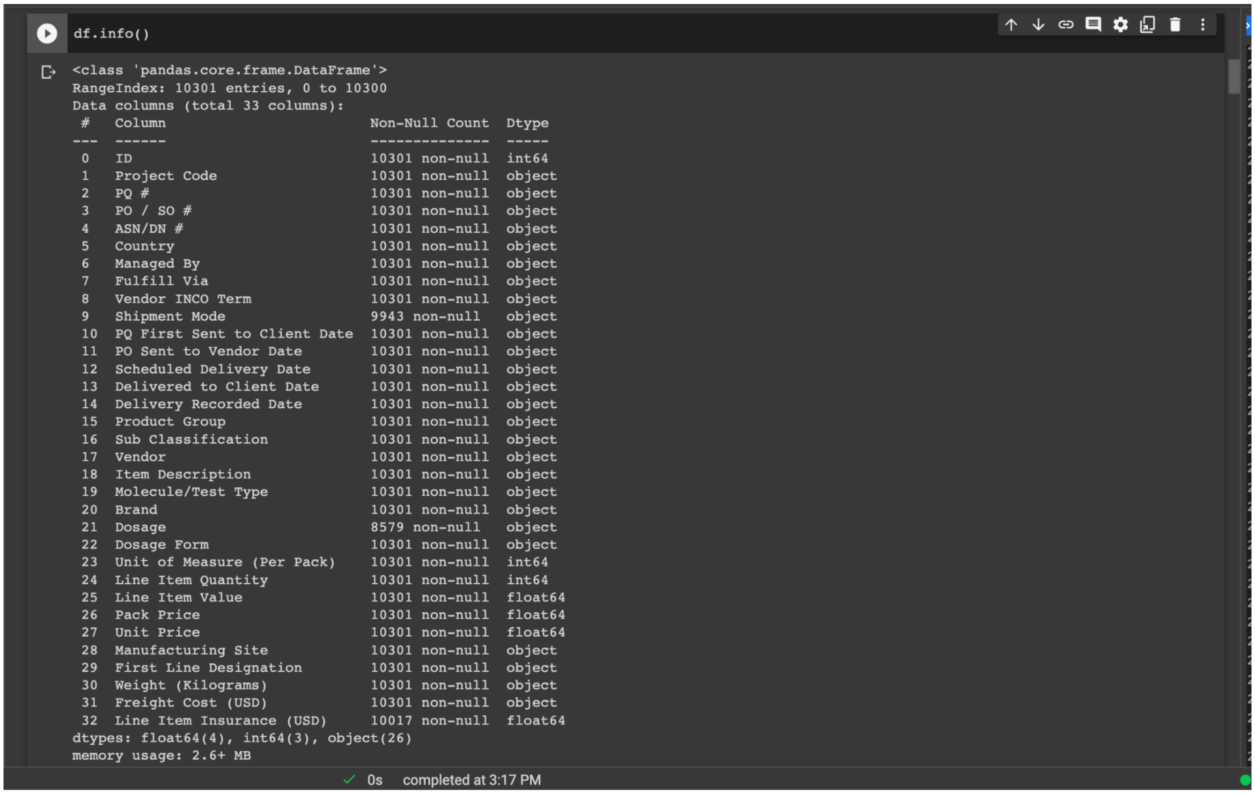
1. Devu Satya Sai Ashish Chandra (801273533)
2. Aarsh Ghotra (801275191)
3. Spandana Pasupuleti (801275952)
4. Mounish Reddy Chintaparthi (801272611)
5. Govinda Satyanarayana Bandaru (801275228)

Data Understanding :

This data set provides supply chain health commodity shipment and pricing data. Specifically, the data set identifies Antiretroviral (ARV) and HIV lab shipments to supported countries. In addition, the data set provides the commodity pricing and associated supply chain expenses necessary to move the commodities to countries for use. The dataset has similar fields to the Global Fund's Price, Quality and Reporting (PQR) data. PEPFAR and the Global Fund represent the two largest procurers of HIV health commodities. This dataset, when analyzed in conjunction with the PQR data, provides a more complete picture of global spending on specific health commodities. The data are particularly valuable for understanding ranges and trends in pricing as well as volumes delivered by country. The US Government believes this data will help stakeholders make better, data-driven decisions. Care should be taken to consider contextual factors when using the database. Conclusions related to costs associated with moving specific line items or products to specific countries and lead times by product/country will not be accurate.

Exploratory Data Analysis :

1. Shape of the data = (10301 , 41) has 10000 rows and 41 columns in the data set.
2. dataFrame.describe
   Shows the top and bottom rows data for all columns.
3. Info of the data frame , gives the information type of the dataset like the column type and how many non null values are present in the column.

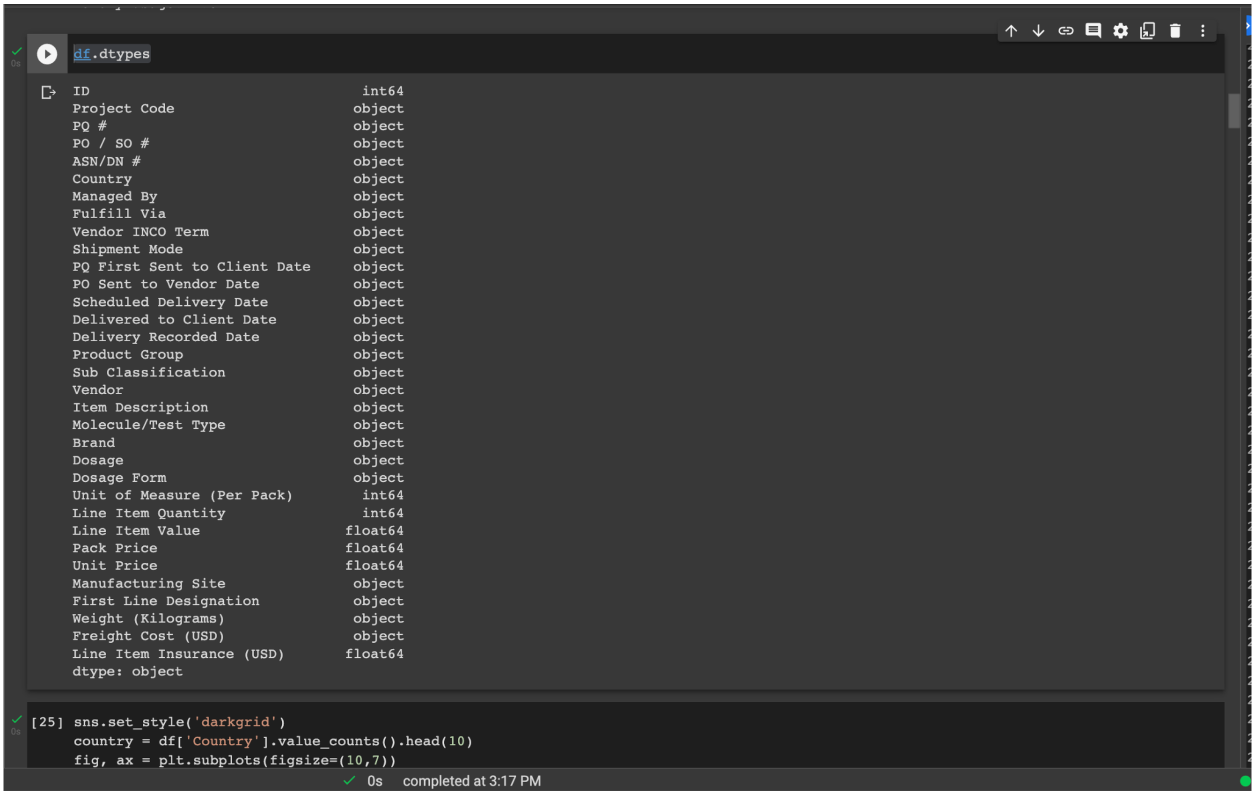


```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10301 entries, 0 to 10300
Data columns (total 33 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   ID                           10301 non-null  int64  
 1   Project Code                 10301 non-null  object 
 2   PQ #                         10301 non-null  object 
 3   PO / SO #                    10301 non-null  object 
 4   ASN/DN #                     10301 non-null  object 
 5   Country                      10301 non-null  object 
 6   Managed By                   10301 non-null  object 
 7   Fulfill Via                  10301 non-null  object 
 8   Vendor INCO Term             10301 non-null  object 
 9   Shipment Mode                9943 non-null   object 
 10  PQ First Sent to Client Date 10301 non-null  object 
 11  PO Sent to Vendor Date       10301 non-null  object 
 12  Scheduled Delivery Date      10301 non-null  object 
 13  Delivered to Client Date     10301 non-null  object 
 14  Delivery Recorded Date       10301 non-null  object 
 15  Product Group                10301 non-null  object 
 16  Sub Classification           10301 non-null  object 
 17  Vendor                       10301 non-null  object 
 18  Item Description             10301 non-null  object 
 19  Molecule/Test Type           10301 non-null  object 
 20  Brand                        10301 non-null  object 
 21  Dosage                       8579 non-null   object 
 22  Dosage Form                  10301 non-null  object 
 23  Unit of Measure (Per Pack)   10301 non-null  int64  
 24  Line Item Quantity           10301 non-null  int64  
 25  Line Item Value              10301 non-null  float64
 26  Pack Price                   10301 non-null  float64
 27  Unit Price                   10301 non-null  float64
 28  Manufacturing Site           10301 non-null  object 
 29  First Line Designation       10301 non-null  object 
 30  Weight (Kilograms)           10301 non-null  object 
 31  Freight Cost (USD)           10301 non-null  object 
 32  Line Item Insurance (USD)    10017 non-null  float64
dtypes: float64(4), int64(3), object(26)
memory usage: 2.6+ MB
```

✓ 0s completed at 3:17 PM

4. Df.dtypes to identify the type of each column



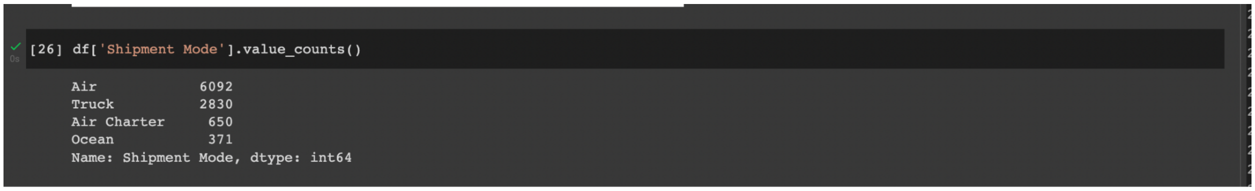
```
df.dtypes
```

```
ID                               int64
Project Code                    object
PQ #                            object
PO / SO #                       object
ASN/DN #                        object
Country                         object
Managed By                      object
Fulfill Via                     object
Vendor INCO Term                object
Shipment Mode                   object
PQ First Sent to Client Date    object
PO Sent to Vendor Date          object
Scheduled Delivery Date         object
Delivered to Client Date        object
Delivery Recorded Date          object
Product Group                   object
Sub Classification              object
Vendor                          object
Item Description                object
Molecule/Test Type              object
Brand                           object
Dosage                          object
Dosage Form                     object
Unit of Measure (Per Pack)       int64
Line Item Quantity               int64
Line Item Value                float64
Pack Price                     float64
Unit Price                     float64
Manufacturing Site              object
First Line Designation          object
Weight (Kilograms)              object
Freight Cost (USD)              object
Line Item Insurance (USD)      float64
dtype: object
```

```
[25] sns.set_style('darkgrid')
     country = df['Country'].value_counts().head(10)
     fig, ax = plt.subplots(figsize=(10,7))
```

✓ 0s completed at 3:17 PM

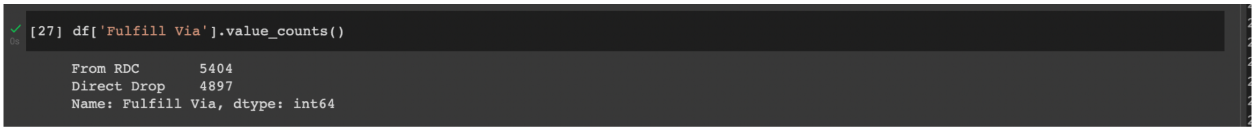
5. Distribution of mode of carriers



```
[26] df['Shipment Mode'].value_counts()

     Air             6092
     Truck           2830
     Air Charter      650
     Ocean            371
     Name: Shipment Mode, dtype: int64
```

6. Type of Deliveries

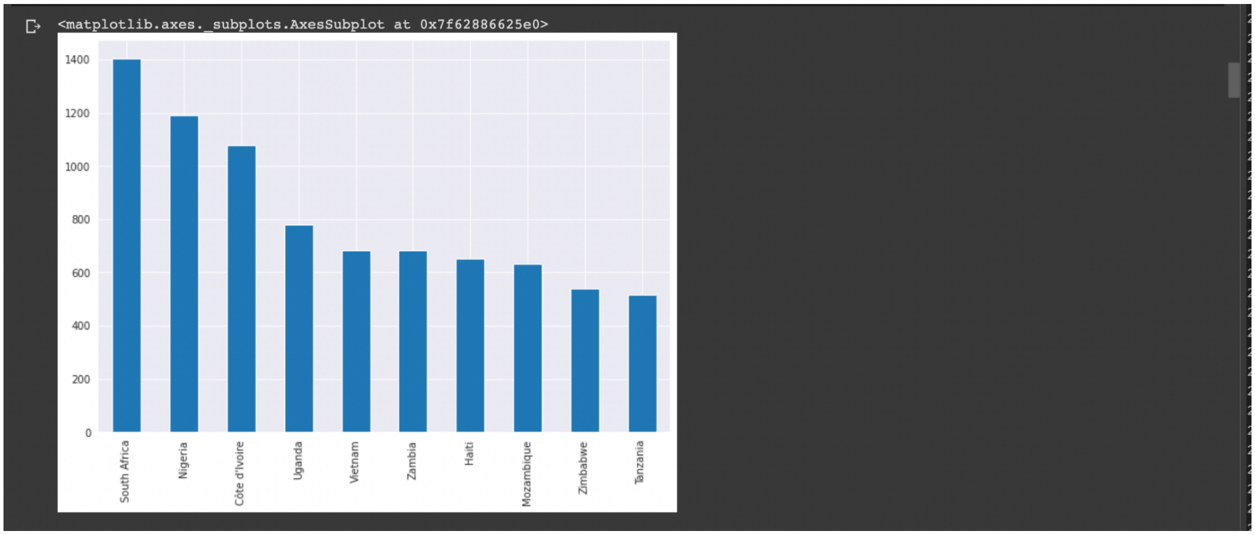


```
[27] df['Fulfill Via'].value_counts()

     From RDC       5404
     Direct Drop    4897
     Name: Fulfill Via, dtype: int64
```

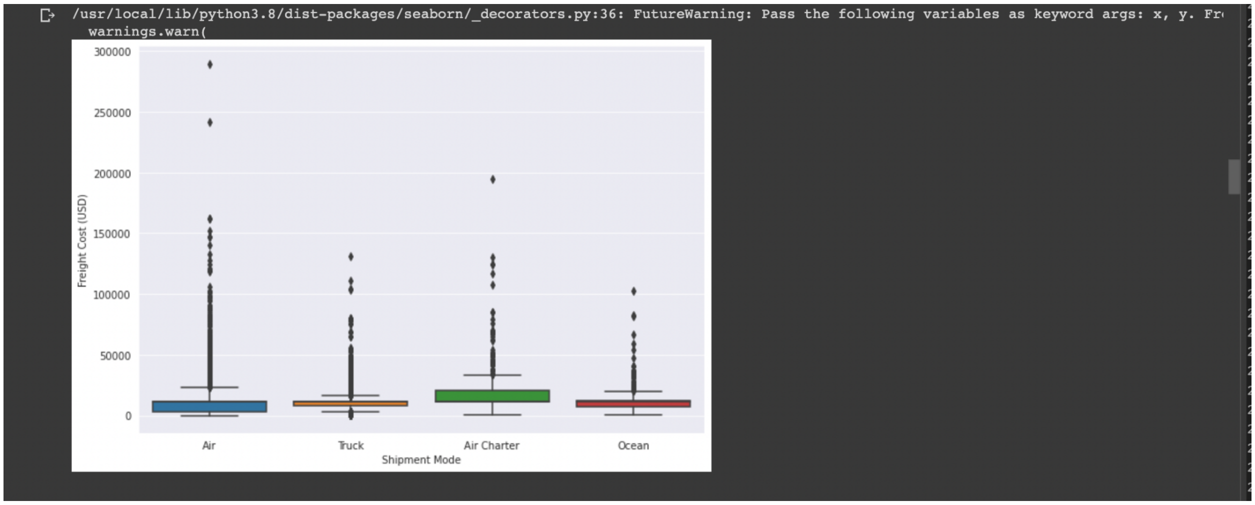
Dashboard : Visualization of Data

Data distributed across the countries



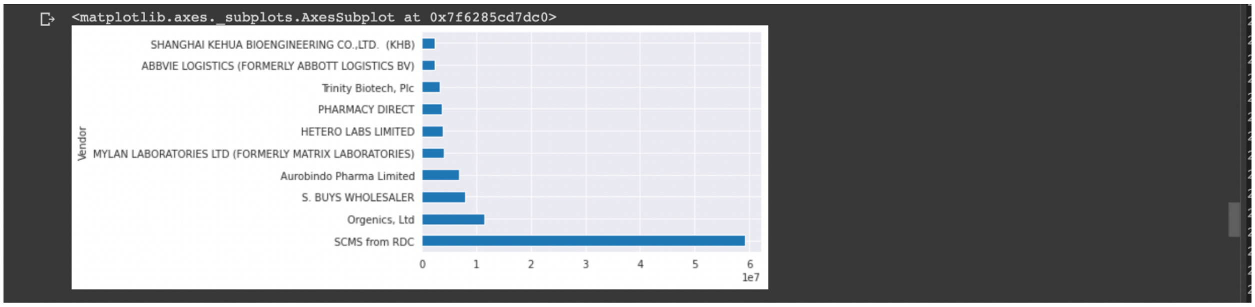
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f62886625e0>
```

ShipmentType Vs the amount of package delivery price.



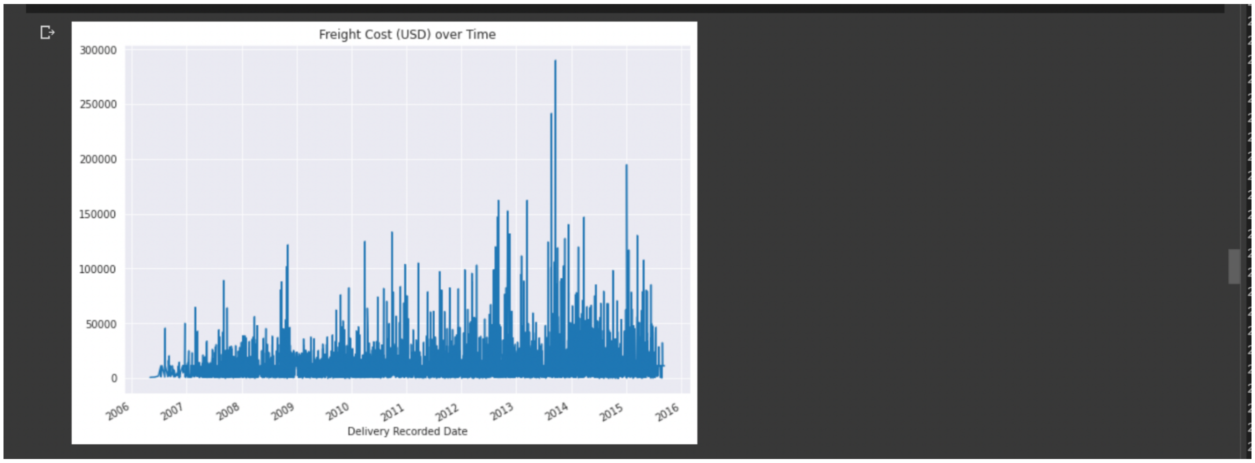
```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. Fr
  warnings.warn(
```

Vendor vs Fright price



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6285cd7dc0>
```

Delivery Record Date vs Fright Price



Data Preparation :

For Data preparation the main modifications are :

1. Convert date columns type to date **Q First Sent to Client Date,PO Sent to Vendor Date,Scheduled Delivery Date,Delivered to Client Date,Delivery Recorded Date**

2. Create feature to be predicted by identifying difference between scheduled and actual delivery

3. Transform Schedule v. Actual column into a categorical int value removing trailing 'days' from values

4. Set all entries with 'Weight Captured Separately' as Null, Replace string values with zero. Update previously transformed zero values with mean value of data

5. Apply same transformations to freight cost feature