Big Data Project: Deliverable 1

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 :

Initially we have 2 data sets provided in the aws **Lab 7 - Supply chain delivery on-time** project. I.e. ShippingLogs.csv (10,000 entries) and ProductDescription.csv (120 entries). Both have a common column called ProductId. We performed a join based on the same column and found that after joining there are 10,000 columns in the data set. After combining both the data sets the following operation are performed to understand the data

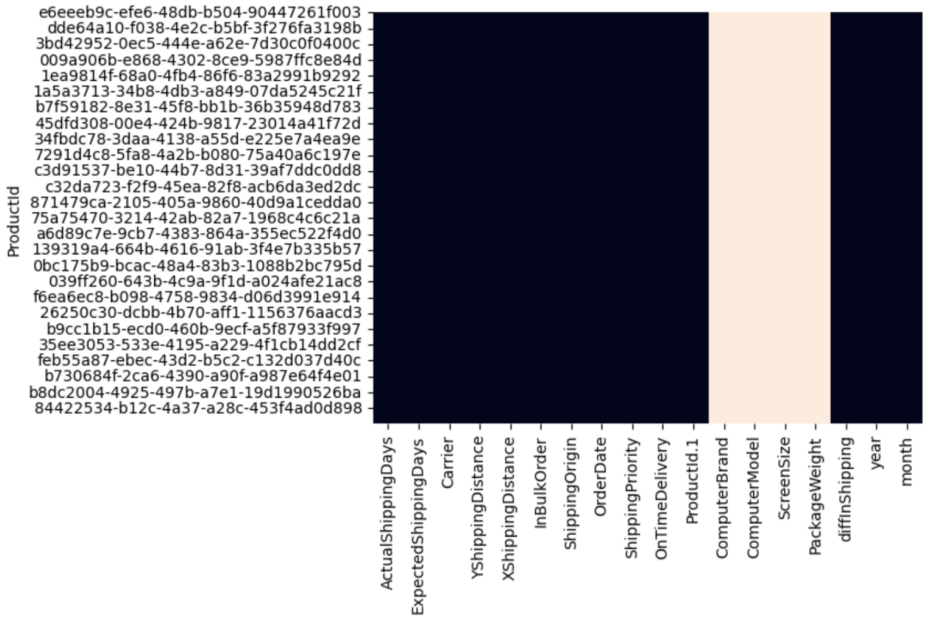
Exploratory Data Analysis :

1. Shape of the data = (10,000 , 15) has 10000 rows and 15 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.

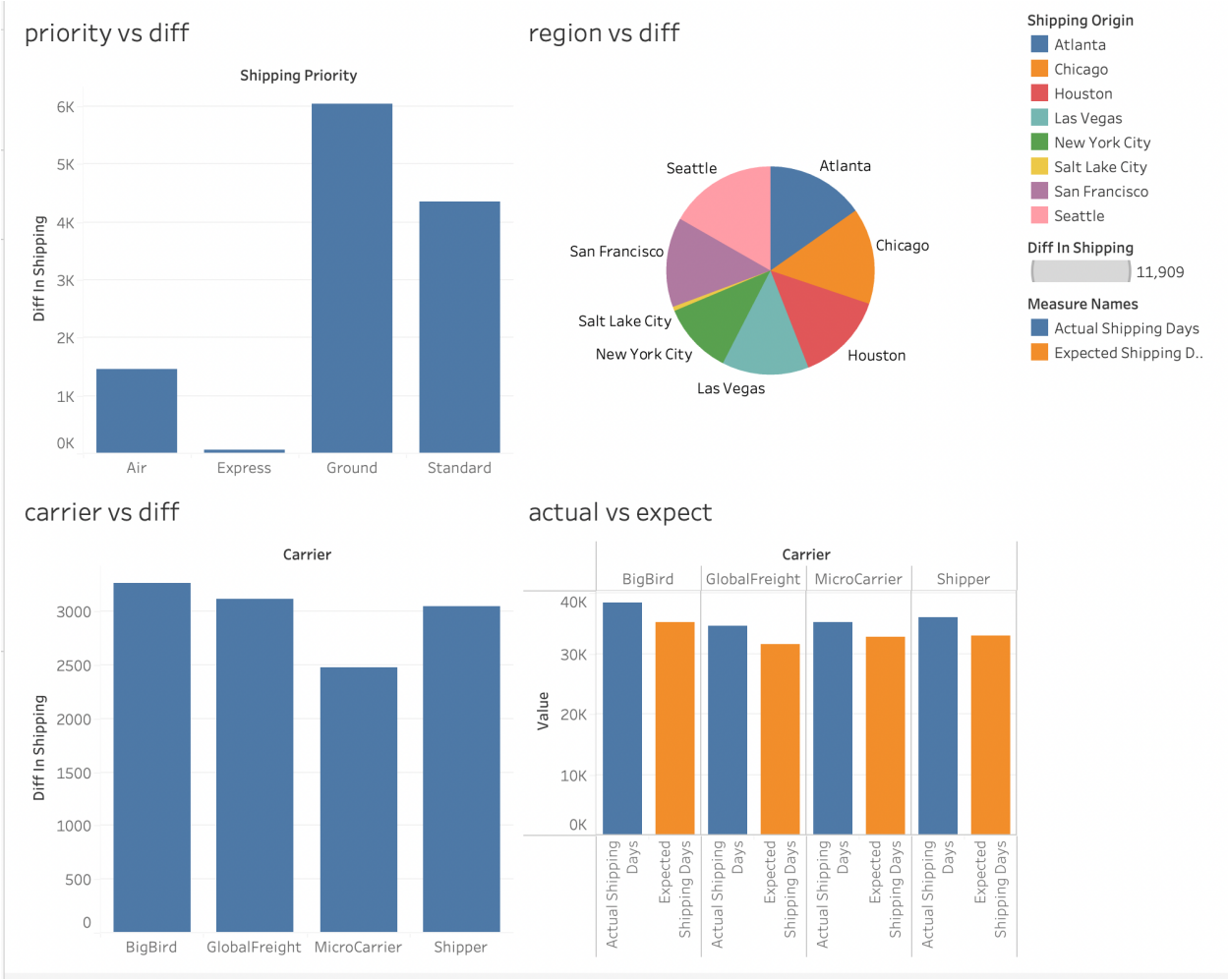
```
In [56]: df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 10000 entries, e6eeeb9c-efe6-48db-b504-90447261f003 to 541e208f-0b96-4797-a866-4784c50ef8e1
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ActualShippingDays    10000 non-null  int64
 1   ExpectedShippingDays  10000 non-null  int64
 2   Carrier               10000 non-null  object
 3   YShippingDistance     10000 non-null  int64
 4   XShippingDistance     10000 non-null  int64
 5   InBulkOrder           10000 non-null  object
 6   ShippingOrigin        10000 non-null  object
 7   OrderDate             10000 non-null  object
 8   ShippingPriority      10000 non-null  object
 9   OnTimeDelivery        10000 non-null  object
 10  ProductId.1           10000 non-null  object
 11  ComputerBrand         0 non-null      object
 12  ComputerModel         0 non-null      object
 13  ScreenSize            0 non-null      float64
 14  PackageWeight         0 non-null      float64
 15  diffInShipping        10000 non-null  int64
dtypes: float64(2), int64(5), object(9)
memory usage: 1.5+ MB
```

4. Dropped duplicate rows if there are any. Fortunately after this step the dimensions are still 10000 X 15. We can state that there are no duplicates.
5. Created a heatmap from SeaBorn to see which column are having null values

```
In [70]: result = sns.heatmap(df.isnull(),cbar=False)
```



Dashboard : (In Tablue)



Data Preparation :

For Data preparation the main modifications are :

1. Joining Two Datasets ShippingLogs.csv and ProductDescription.csv based on ProductId
2. Extract column called DifferenceInShipping days which essentially tells us how many extra days did the package take to arrive at the destination (Actual-expected shipping days.)
3. Extracted years and months to a separate columns to perform additional analysis for the delivery delay trend
4. After performing the above steps, we extracted the data into a new csv file and presented the above visualizations.