UCD
DUBLIN

COMP47350 - Data Analytics

Data Quality Report

Assignment 1

Stephen Moles: 11371826

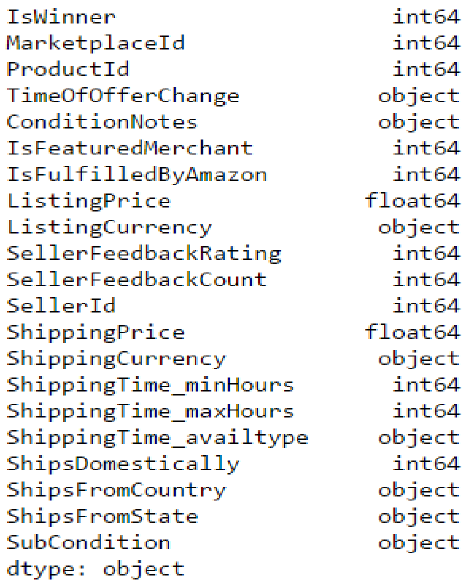
Contents

1.0 Statistical Analysis

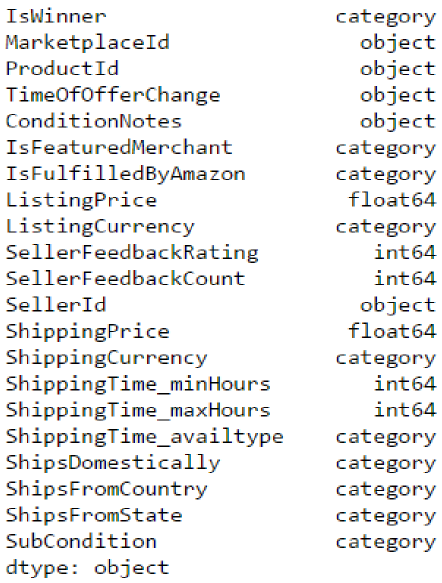
This document will outline the initial quality of the presented data. The report will evaluate and summarise the features present in the current data, outlining areas where initial changes have been made. This initial changes will focus on; correcting or establishing appropriate feature types, and the dropping of constant columns or duplicate rows. The result will be more accurate tables and visualisations of the Continuous and Categorical features, which will provide some initial insights about the data, and aid in the preparation of the Data Quality Plan.

1.1 Feature Types

Below is a list of the initial features, both before (left-hand side) and after (right-hand side) changes have been made to their data types:



```
IsWinner                       int64
MarketplaceId                  int64
ProductId                      int64
TimeOfOfferChange             object
ConditionNotes                object
IsFeaturedMerchant             int64
IsFulfilledByAmazon            int64
ListingPrice                 float64
ListingCurrency               object
SellerFeedbackRating           int64
SellerFeedbackCount            int64
SellerId                       int64
ShippingPrice                float64
ShippingCurrency              object
ShippingTime_minHours          int64
ShippingTime_maxHours          int64
ShippingTime_availtype        object
ShipsDomestically              int64
ShipsFromCountry              object
ShipsFromState                object
SubCondition                  object
dtype: object
```



```
IsWinner                   category
MarketplaceId                object
ProductId                    object
TimeOfOfferChange            object
ConditionNotes               object
IsFeaturedMerchant         category
IsFulfilledByAmazon        category
ListingPrice                float64
ListingCurrency            category
SellerFeedbackRating          int64
SellerFeedbackCount           int64
SellerId                     object
ShippingPrice               float64
ShippingCurrency           category
ShippingTime_minHours         int64
ShippingTime_maxHours         int64
ShippingTime_availtype     category
ShipsDomestically          category
ShipsFromCountry           category
ShipsFromState             category
SubCondition               category
dtype: object
```

The following features had their data types changed:

- IsWinner: From Continous to Categorical
- MarketplaceId: From Continous to Object
- ProductId: From Continous to Object
- IsFeaturedMerchant: From Continous to Categorical
- IsFulfilledByAmazon: From Continous to Categorical
- ListingCurrency: From Object to Categorical
- SellerID: From Continous to Object
- ShippingCurrency: From Object to Categorical
- ShippingTime_availtype: From Object to Categorical
- ShipsDomestically: From Continous to Categorical
- ShipsFromCountry: From Object to Categorical
- ShipsFromState: From Object to Categorical
- Subcondition: From Object to Categorical

There were various motivations for changing these features and they will be categorised in the following order:

The features; *IsWinner, IsFeaturedMerchant, IsFulfilledByAmazon, ShipsDomestically* should be grouped together. Although these features are represented by integers, their value should not be considered to be numerical or continuous. The numbering used represents a binary Yes or No value, and for this reason it is more accurate to change these types to Categorical.

The features; *MarketplaceId, ProductId, and SellerId* should be grouped together. Although these features have been initally marked as Categorical features, they shoul be changed to Objects. The reason behind is this decision is that ID's do not represent Continuous or Categorical information. Therefore, for these features to be included in these types, it may skew or corrupt potential statistics or visualisations about the data. For this reason they hace been altered to Objects. For example, if the length of unique values for these three features is checked, it reveals that the for one unique market place, there are 187 unique seller ID's, selling 307 unique products.

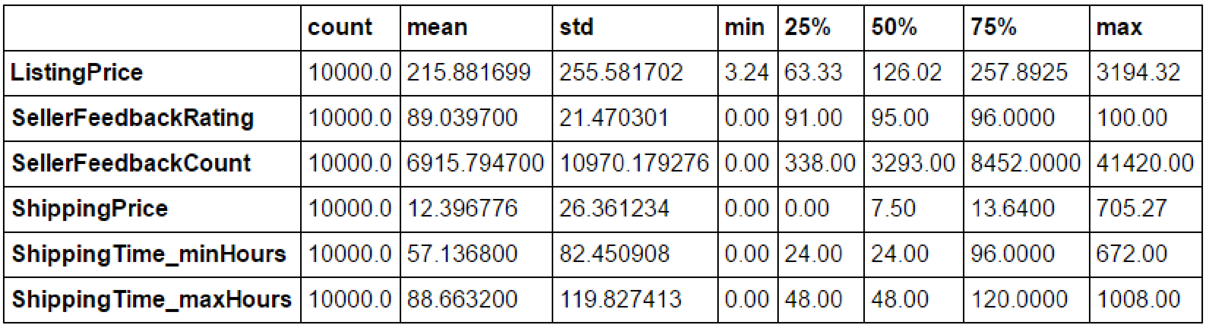
The features; *ListingCurrency, ShippingCurrency, ShippingTime_availtype, ShipsFromCountry, ShipsFromState,* and *Subcondition* should be grouped together. These features have initally been listed as Objects, but should be changed to Categorical, as they represent important statisitcal information which can be interpreted by tables and visualisations.

1.2 Drop Duplicates

The following tables repesent the statisical Continuous data before (top table) and after (bottom table) constant columns and duplicate rows have been dropped:

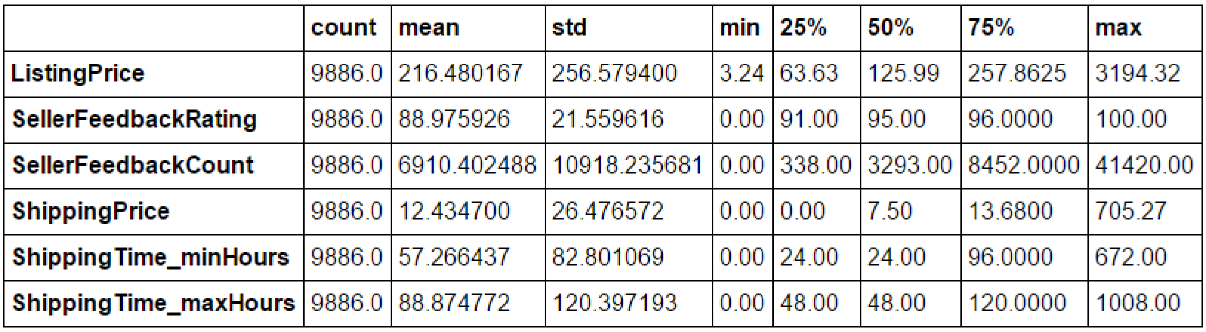
1.2.1 Table of Continuous Features

Before:



	count	mean	std	min	25%	50%	75%	max
ListingPrice	10000.0	215.881699	255.581702	3.24	63.33	126.02	257.8925	3194.32
SellerFeedbackRating	10000.0	89.039700	21.470301	0.00	91.00	95.00	96.0000	100.00
SellerFeedbackCount	10000.0	6915.794700	10970.179276	0.00	338.00	3293.00	8452.0000	41420.00
ShippingPrice	10000.0	12.396776	26.361234	0.00	0.00	7.50	13.6400	705.27
ShippingTime_minHours	10000.0	57.136800	82.450908	0.00	24.00	24.00	96.0000	672.00
ShippingTime_maxHours	10000.0	88.663200	119.827413	0.00	48.00	48.00	120.0000	1008.00

After:



	count	mean	std	min	25%	50%	75%	max
ListingPrice	9886.0	216.480167	256.579400	3.24	63.63	125.99	257.8625	3194.32
SellerFeedbackRating	9886.0	88.975926	21.559616	0.00	91.00	95.00	96.0000	100.00
SellerFeedbackCount	9886.0	6910.402488	10918.235681	0.00	338.00	3293.00	8452.0000	41420.00
ShippingPrice	9886.0	12.434700	26.476572	0.00	0.00	7.50	13.6800	705.27
ShippingTime_minHours	9886.0	57.266437	82.801069	0.00	24.00	24.00	96.0000	672.00
ShippingTime_maxHours	9886.0	88.874772	120.397193	0.00	48.00	48.00	120.0000	1008.00

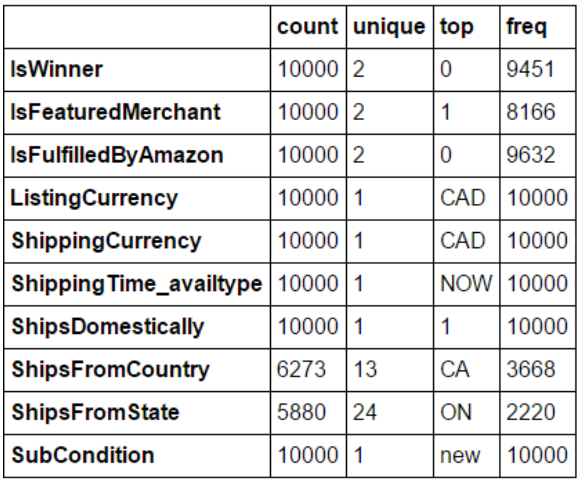
1.2.2 Insights and Data Processing – Continuous Features

Firstly, in relation to constant columns, it is clear from the two table there within the data, there were no columns which contained constant values which merited being dropped. However, in terms of rows, there has been a reduction in rows from the initall 10,000 to 9886. The effect of this has been noted within the statistics, with a reduction related to count and median ListingPrice value, and a marginal increase associated to the mean, with the exception of SellerFeedbackRating and SellerFeedbackCount, and the standard deviation, with the exception of SellerFeedbackRating.

1.2.3 Table of Categorical Features

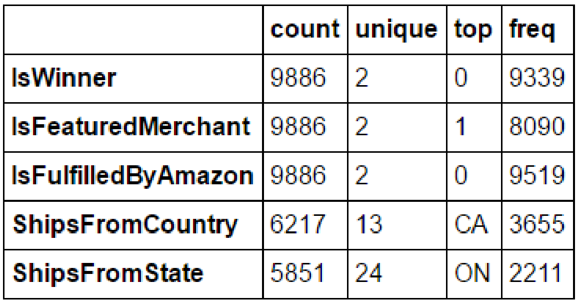
The following tables repesent the statisical Categorical data before (top table) and after (bottom table) constant columns and duplicate rows have been dropped:

Before:



	count	unique	top	freq
IsWinner	10000	2	0	9451
IsFeaturedMerchant	10000	2	1	8166
IsFulfilledByAmazon	10000	2	0	9632
ListingCurrency	10000	1	CAD	10000
ShippingCurrency	10000	1	CAD	10000
ShippingTime_availtype	10000	1	NOW	10000
ShipsDomestically	10000	1	1	10000
ShipsFromCountry	6273	13	CA	3668
ShipsFromState	5880	24	ON	2220
SubCondition	10000	1	new	10000

After:



	count	unique	top	freq
IsWinner	9886	2	0	9339
IsFeaturedMerchant	9886	2	1	8090
IsFulfilledByAmazon	9886	2	0	9519
ShipsFromCountry	6217	13	CA	3655
ShipsFromState	5851	24	ON	2211

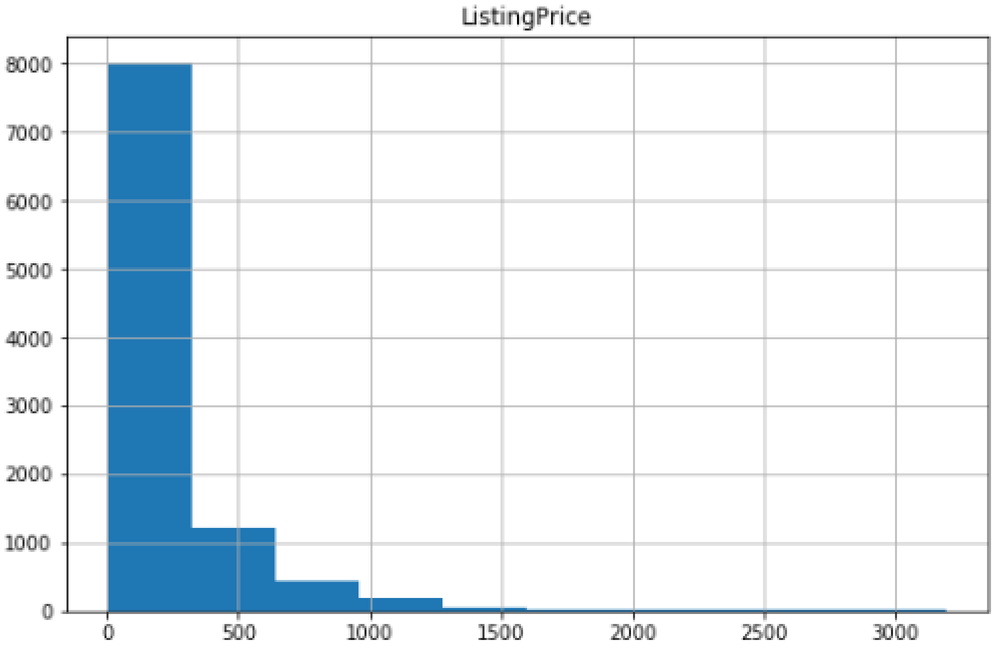
1.2.4 Insights and Data Processing – Categorical Features

Firstly in relation to constant columns, it is significant to note that there has been a reduction in columns from 10 to 5. The columns were deleted on the basis that the values were constant throughout, which would offer very little to the analysis of the data by obscuring the results. The deletion of duplicate rows has also effected the data by a reduction in the count of rows and the frequency of these features in the data.

2.0 Analysis of Visualisations

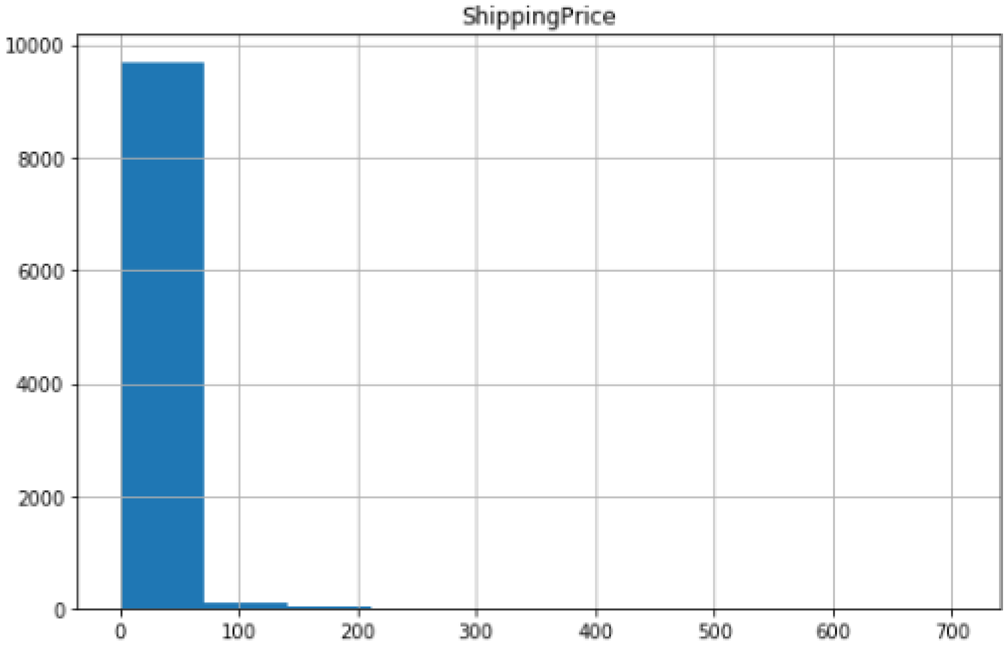
The following section will provide visualisations of Histograms and Boxplots, representing Continuous Data, and of Bar Plots, which will represent the Categorical data, after the aforementioned changes have been made to the statistical data.

2.1 Histograms – Continuous Features



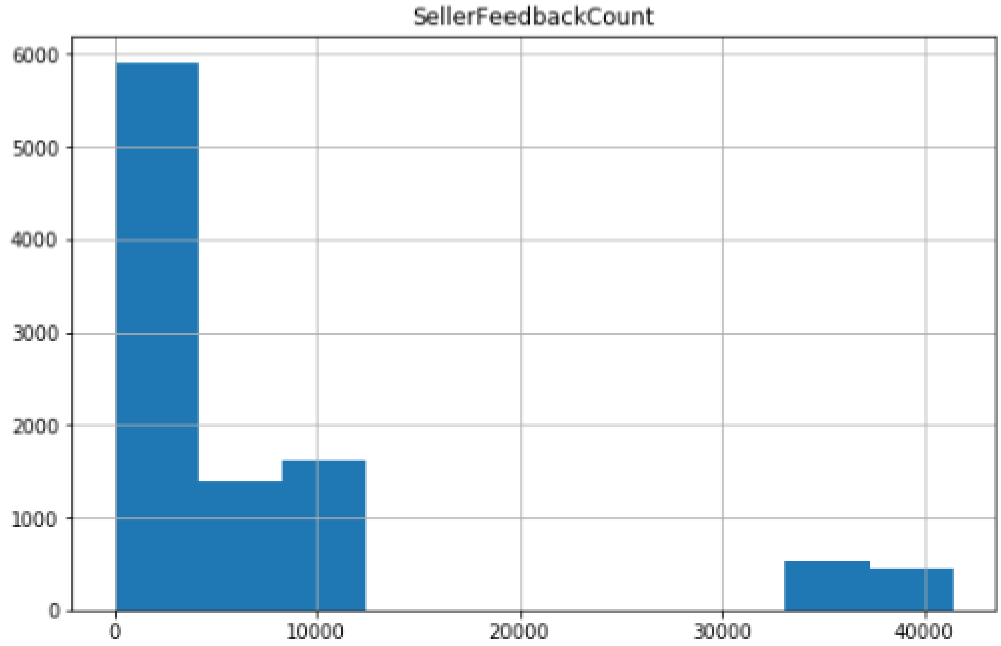
2.1.1 Listing Price: Initial Analysis

The *ListingPrice* visualisation indicates that the majority of sellers, regardless of having been recognised as a winner, list their prices between roughly 0 and 400. A smaller sample indicate higher prices, and the Histogram suggest that there may be outliers. These outliers will be more visible in Boxplot diagrams at a later stage, and potential strategies to deal with these will be address in the Data Quality Plan.



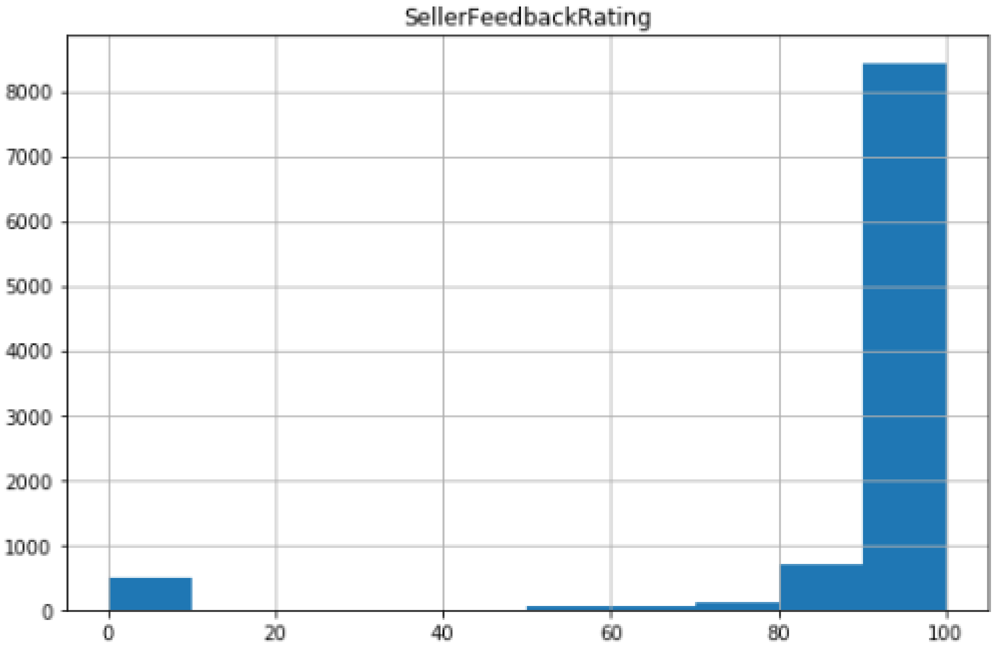
2.1.2 Shipping Price: Initial Analysis

Initial analysis of the above Histogram suggest that most sellers either do not charge a shipping price, or charge no more than about 60 for shipping. There are some notable exceptions, where shipping can cost between 60 and 210. Further analysis will be provided in the boxplot, but further analysis will be required later to address how many do not charge a shipping price.



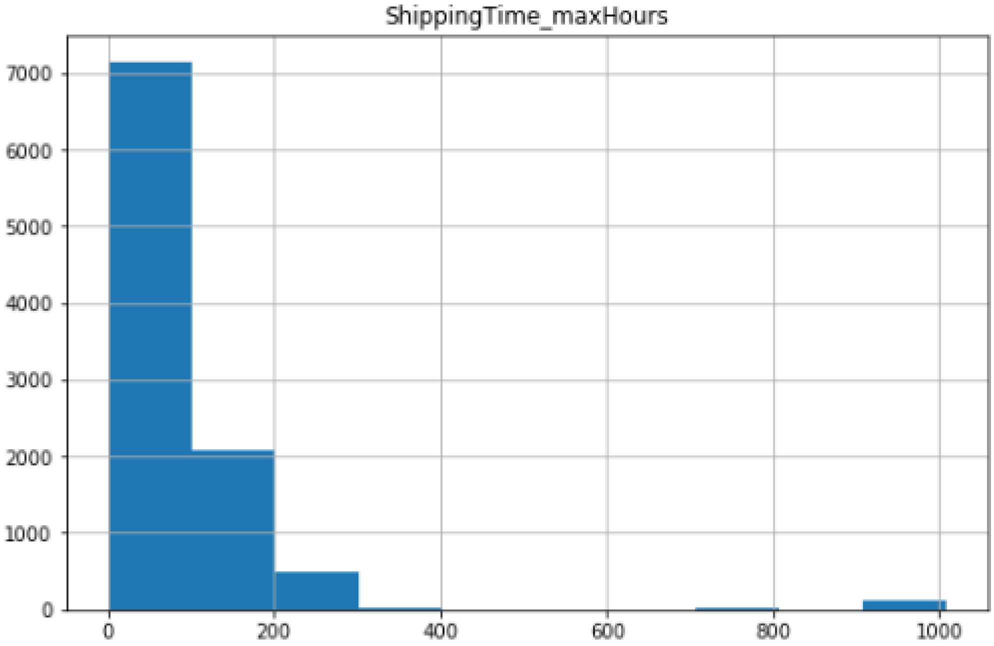
2.1.3 Seller Feedback Count: Initial Analysis

This section offers statistics that indicates that for a large number, almost 6000 of the rows did not include any feedback for the seller. Perhaps this will have contributed to the high number of *IsWinner* returning false, and should be addressed. If this turns out to be the case, it may explain the large gap to where the SellerFeedbackCount exceeds 40,000. Perhaps these attribute to the result of *IsWinner* returning as true, or 1.



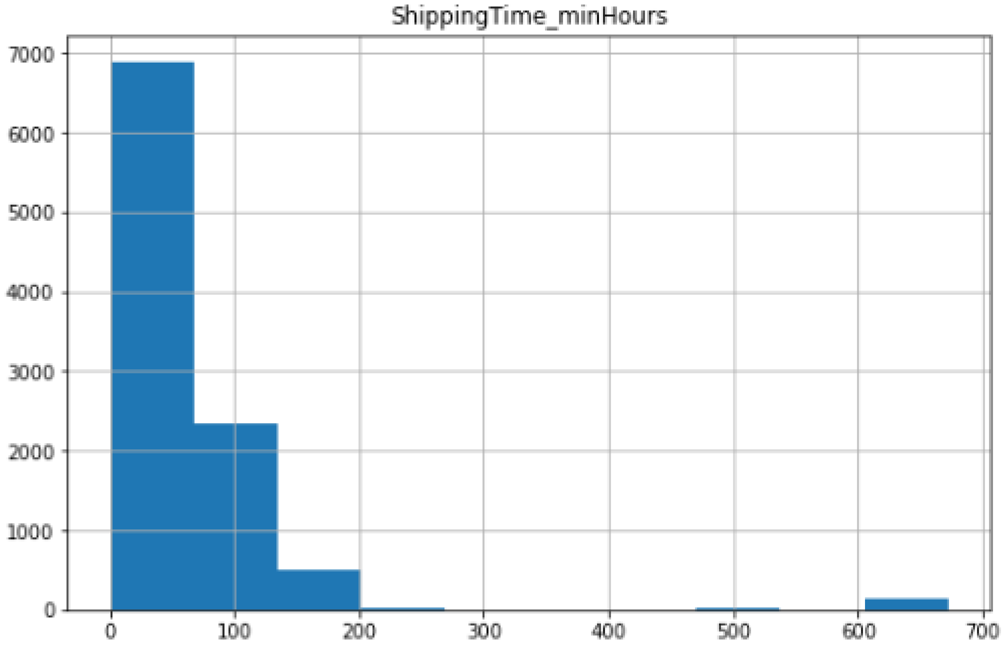
2.1.4 Seller Feedback Rating: Initial Analysis

The above visualisation indicates that a large number of Seller Feedback, more than 8,000 returned the highest rating. Amounts lower than 1,000 indicate that some feedback ratings were rated 0. It should be noted whether these are due to null values where feedback isn't given, whether there is an element of human error, or if 0 is assigned as the default value should no amount be entered.



2.1.5 Shipping Time – Max Hours: Initial Analysis

The above Histogram indicates that the majority of products have a maximum shipping time of 100 hours, or about 4 days. Others can take from 4 to 16 days, which could be due to the difference needed to travel for delivery. There are outliers at the upper quartile, which could either be due to international delivery, or perhaps attributed to human error.

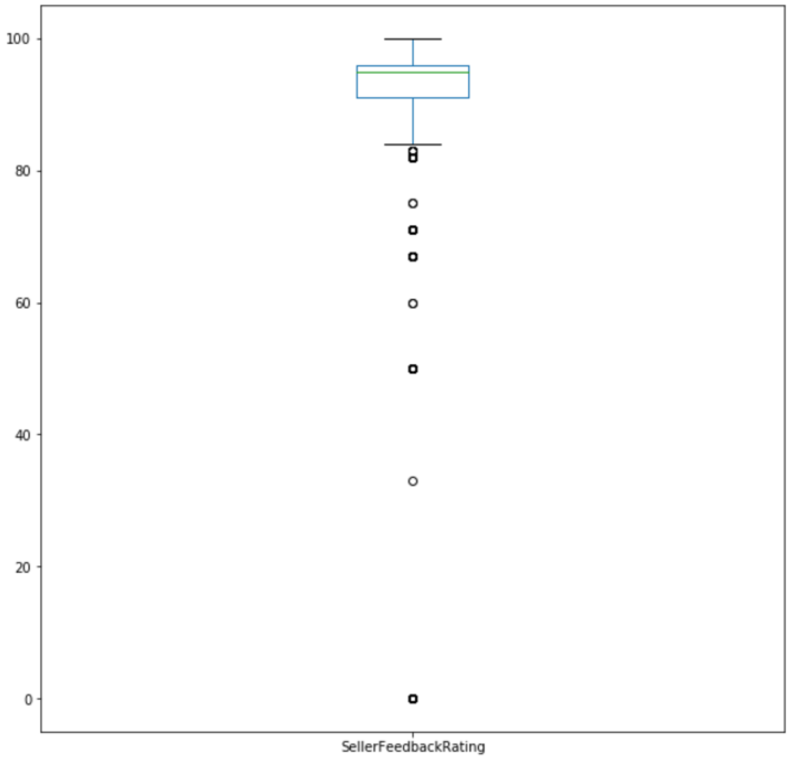


2.1.6 Shipping Time – Min Hours: Initial Analysis

For most products, almost 7000 thousand rows indicate that at least 3 days or around 60 hours are required at least for shipping. There are other outliers which indicate that at least

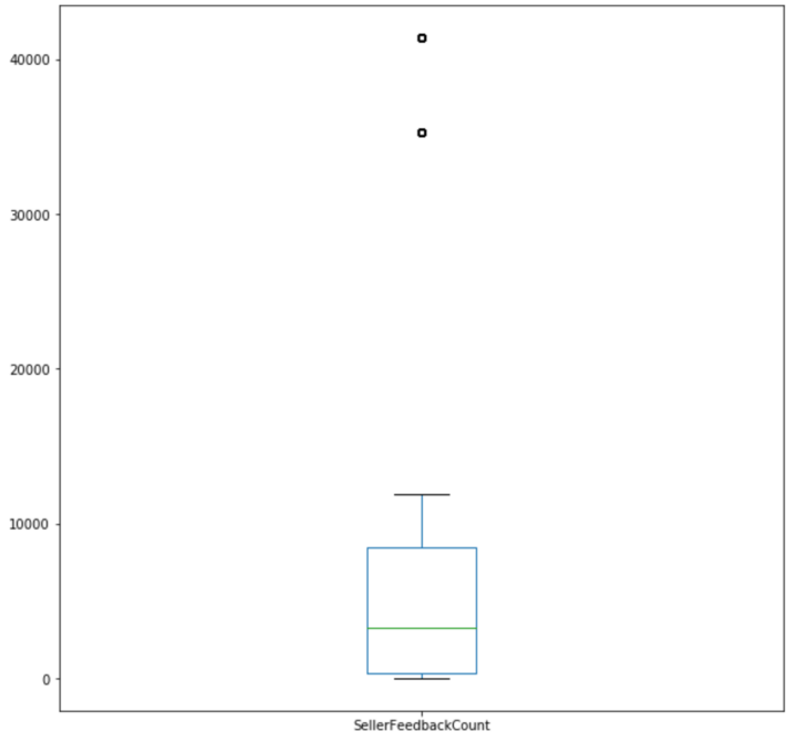
600 hours should be expected for shipping. It’s unclear if this can be attributed to human error, or whether this may represent international deliveries.

2.2 Boxplots – Continuous Features



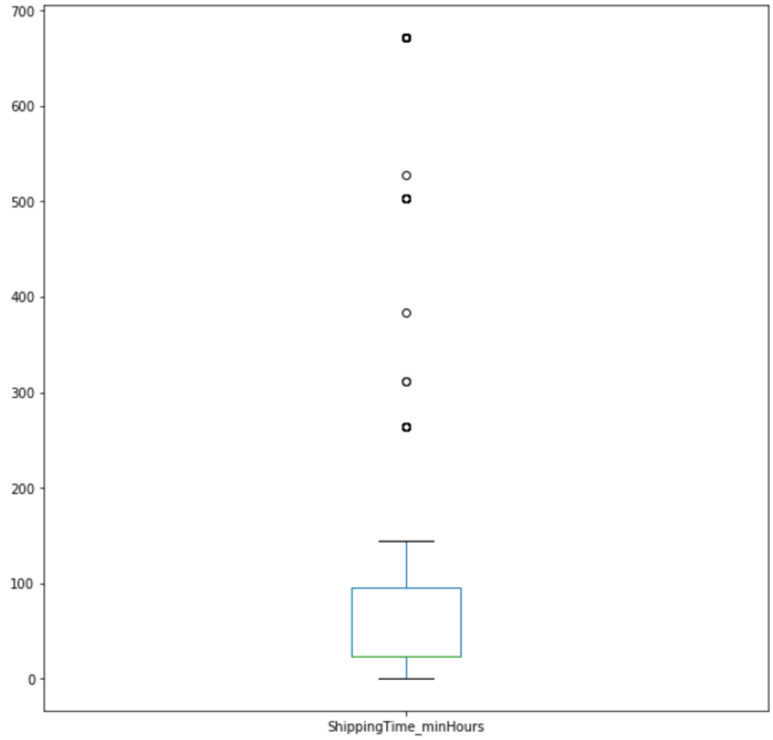
2.2.1 Seller Feedback Rating: Initial Analysis

This plot highlights the amount of outliers beyond the minimum threshold. It shows that a small portion of the rating as are in the third quartile above the median. Most ratings fall below the median and represented in the first quartile.



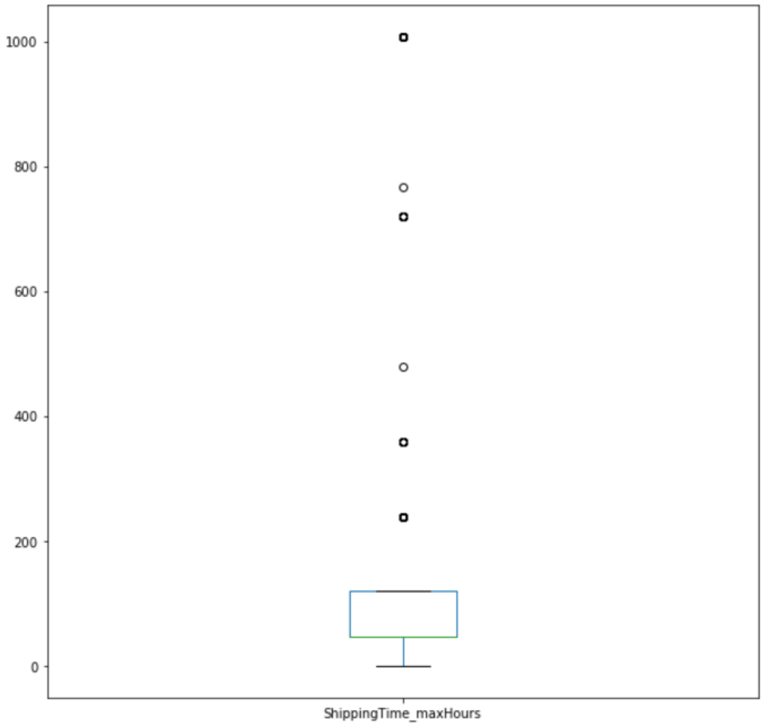
2.2.2 Seller Feedback Count: Initial Analysis

The data suggests that the median lies at a count of around 4000. There is a large portion of this count located in the upper quartile, with some outliers plotted beyond the maximum indicated threshold. A lower number of the count is represented in the lower quartile.



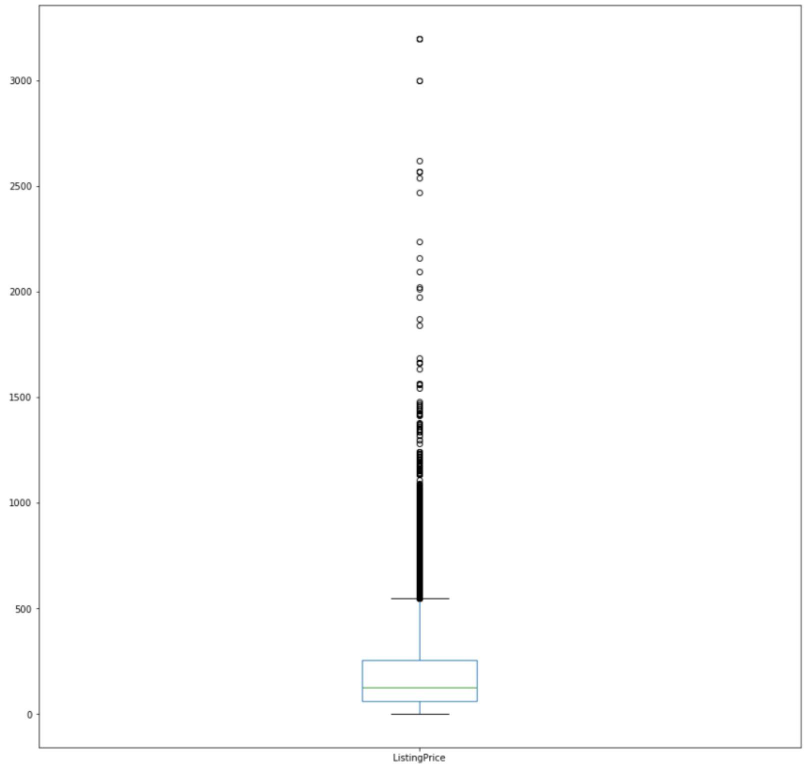
2.2.3 Shipping Time – Min Hours: Initial Analysis

The boxplot's data suggests that the median minimum hours is located at the intended lower quartile section of the box. This could be due the significant number of outliers that fall well above the maximum threshold.



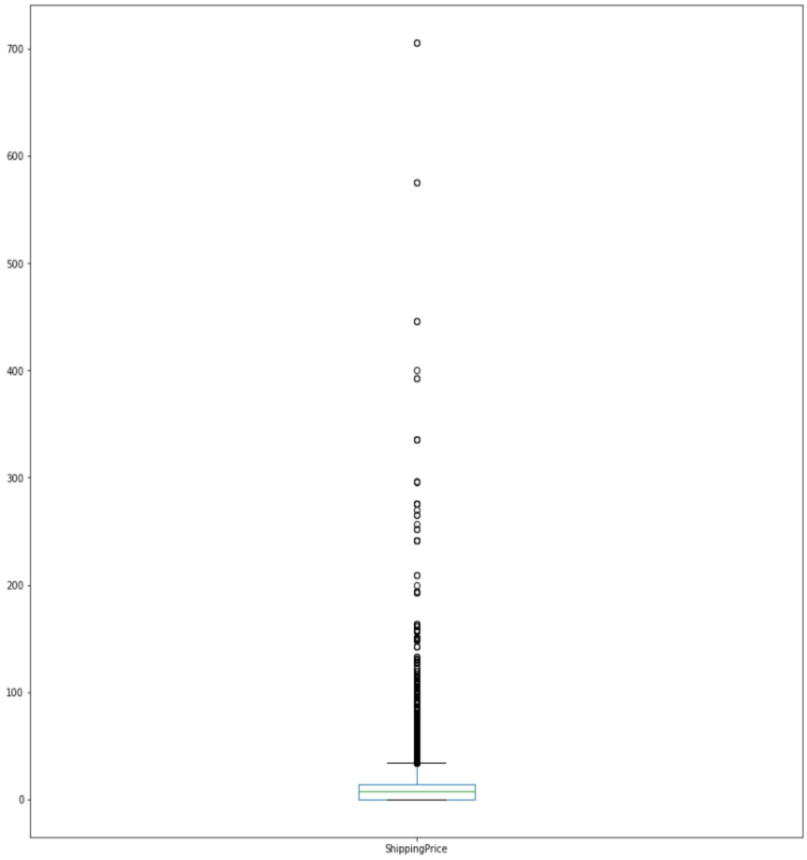
2.2.4 Shipping Time – Max Hours: Initial Analysis

Similarly in the above boxplot, the median is located in the lower quartile of the box. The maximum threshold occurs immediately after the upper quartile section. It could be that the outliers above are skewing the results of this feature. Measures to address outliers will be discussed within the Data Quality Plan.



2.2.5 Listing Price: Initial Analysis

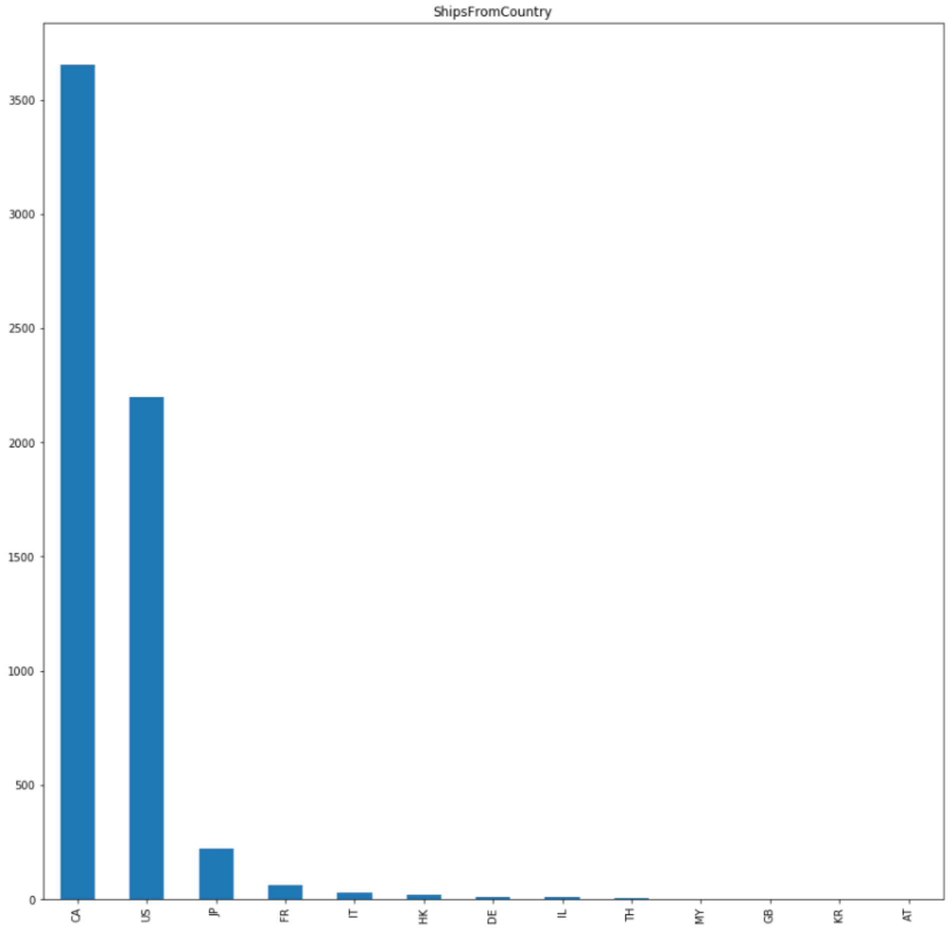
The *ListingPrice* contains a significant number of outliers. A majority of products fall above the median, with the remainder falling with the lower quartile range.



2.2.6 Shipping Price: Initial Analysis

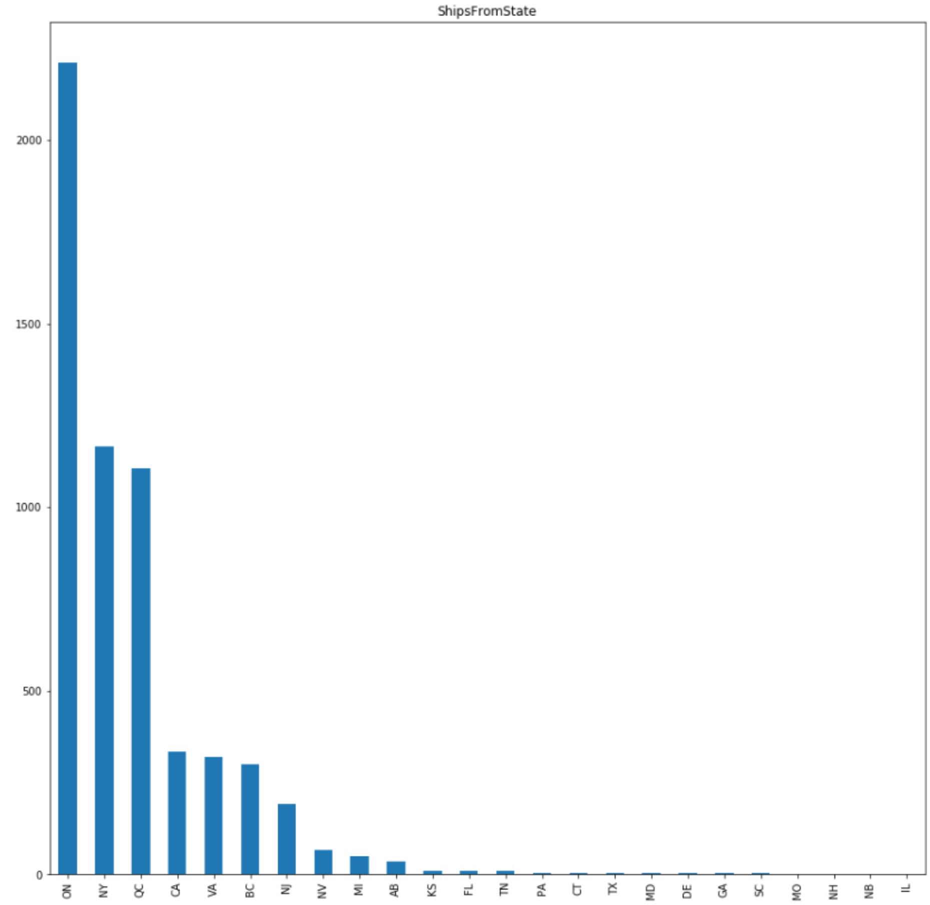
Similarly, the feature *ShippingPrice* contains a significant number of outliers which are higher than the maximum threshold. The size of the upper and lower quartiles are almost equal in size, with the median falling just above 0.

2.3 Bar Charts – Categorical Features



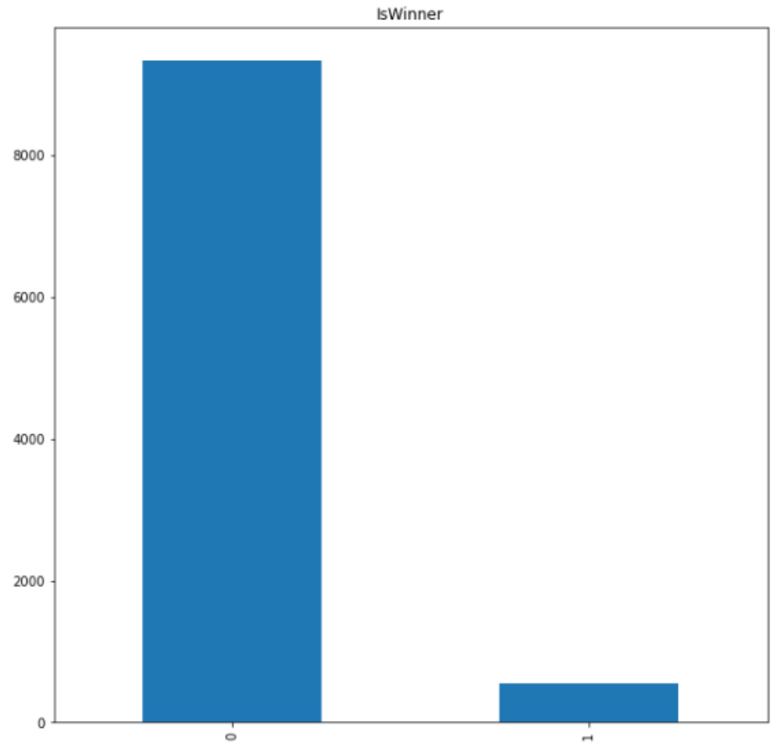
2.3.1 Ships from Country: Initial Analysis

The above chart may clarify why some of the shipping hours appear to lie outside of the expected median and quartile ranges. As is indicated, the largest countries that ship the products are Canada, and the United States. The shipping hours from these two countries to the international market may manipulate the shipping time data.



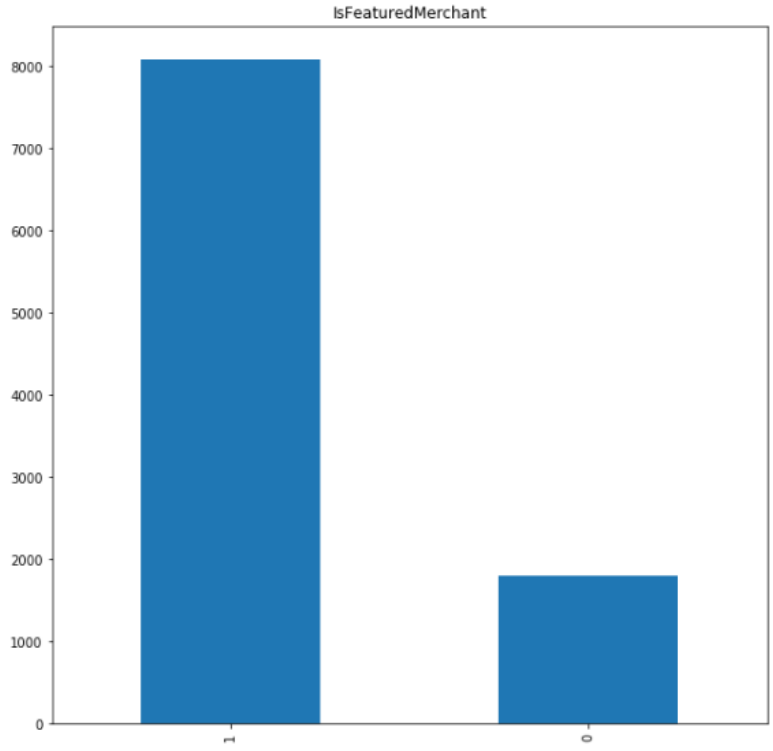
2.3.2 Ships from State: Initial Analysis

The data provided establishes Canadian provinces such as Ontario, Quebec, and British Columbia, and the American state of New York as the largest shipping states and provinces. The data is bit unclear as in terms of differentiating between states and provinces. As there is more data available about the three largest shipping territories, perhaps location and shipping capabilities will play an important role in deciding whether a seller *IsWinner*.



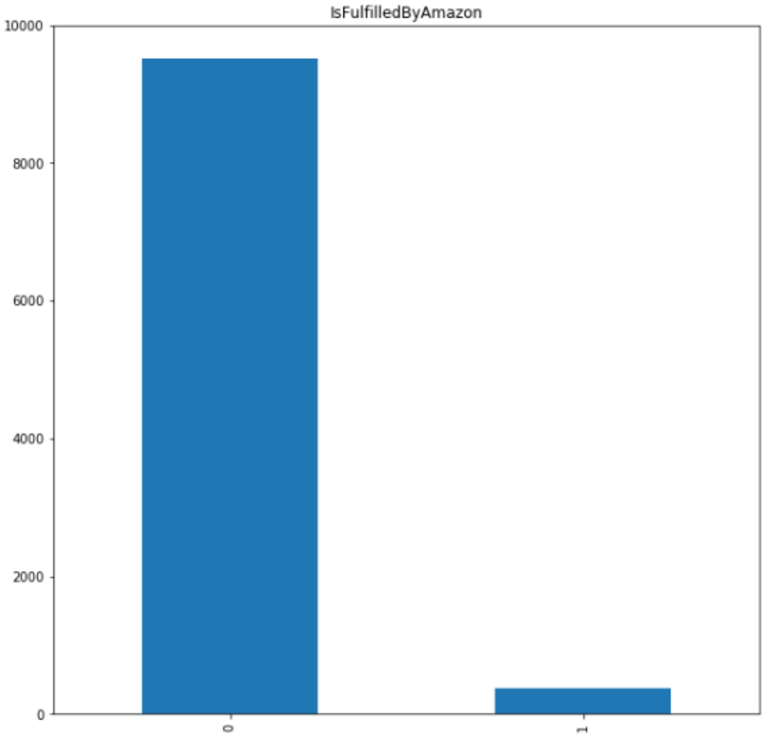
2.3.3 Is Winner: Initial Analysis

Interestingly, the majority of the data provided accounts for those that are not categorised as a Winner. A small margin, some 549, account for those that have been selected as *IsWinner* by Amazon. Analysing the feature values that these 549 rows share, may indicate some of the characteristics that are required to be selected.



2.3.4 Is Featured Merchant: Initial Analysis

The above bar chart illustrates that the majority of rows indicate that most products that are being sold, are being sold by Featured Merchants. It could be important to see does this majority correlate with those who have been marked as Winners, or if it represents those who have not.



2.3.5 Is Fulfilled By Amazon: Initial Analysis

The majority of products being sold by sellers are not fulfilled by Amazon. Of the 9886 rows, 9519 of these are not fulfilled by Amazon, as represented by 0 which is attributed to the Boolean False. The remainder, 367 are currently being fulfilled by Amazon.

3.0 Conclusion

The above visualisations aim to provide an understanding of the current health of the data, and what it may indicate are potentially insightful features. The above data currently does not provide any concrete indications, and therefore its quality should be accurately assessed. There are a significant number of outliers for many of the features, and certain features currently contain a significant amount of null values. Further in depth analysis of potential Data Quality Issues should be outlined and later addressed within a Data Quality Plan.