DATA QUALITY REPORT

HOMEWORK 1

Giuseppe Benanti –ID 07266120

INTRODUCTION

Extensive explanation of the different CRISP-DM steps for this predictive analytics project are given in the correlated Jupyter notebook (*Homework1_Comp47350_Benanti_07266120.ipynb*).

Here we summarize in tables and graphs the results from the **Data Exploration** step of the CRISP-DM process deployed in Anaconda-Jupyter environment using Python 3.6 on the dataset stored in the file *amazon-offers-10k-samples-raw.csv.*

Below, Tables 1 to 4 list descriptive statistics results from the original dataset on offers from vendors on Amazon.

The original dataset is comprised of a totality of 10000 entries (rows), one per offer, and 21 columns, with 20 descriptive features and 1 target feature, and a numeric column for row number.

Each product can be offered by many vendors, but only one of these vendors (per product) will become a 'Winner' vendor and listed at the top of the user research page. The binary categorical target feature 'IsWinner' expresses this concept, 1 for winner, 0 for looser.

The predictive model will attempt to predict the target feature based on a selection of the descriptive features data.

1 - CONTINUOUS FEATURES

Table 1. Descriptive statistics results for continuous (numerical) features. Dataset: *amazon-offers-10k-samples-raw.csv.*

Feature	Count	%Miss.	Card.	Min.	1stQRT	Mean	Median	3rdQRT	Max.	StDev.
MarketplaceId	10000	0.0	1	1.15646e+18	1.15646e+18	1.15646000000000553e+18	1.15646e+18	1.15646e+18	1.15646e+18	55298.7
ProductId	10000	0.0	307	-9.1726e+18	-4.35648e+18	5.534224590000009e+16	2.47757e+16	5.0561e+18	9.20715e+18	5.308389425152429e+18
ListingPrice	10000	0.0	4976	3.2	63.3	215.8	126.0	257.9	3194.3	255.6
SellerFeedbackRating	10000	0.0	25	0.0	91.0	89.0397	95.0	96.0	100.0	21.5
SellerFeedbackCount	10000	0.0	123	0.0	338.0	6915.7	3293.0	8452.0	41420.0	10970.2
SellerId	10000	0.0	187	-9.13612e+18	-2.57228e+18	9.780680312000013e+17	1.16312e+18	5.47794e+18	9.17405e+18	4.952648039850679e+18
ShippingPrice	10000	0.0	1104	0.0	0.0	12.40	7.5	13.6	705.3	26.3
ShippingTime_minHours	10000	0.0	12	0.0	24.0	57.2	24.0	96.0	672.0	82.4
ShippingTime_maxHours	10000	0.0	11	0.0	48.0	88.7	48.0	120.0	1008.0	119.8
ShipsDomestically	10000	0.0	1	1.0	1.0	1.0	1.0	1.0	1.0	0.0

Table 2. Descriptive statistics results for continuous (numerical) features. Revised table after removal of duplicates and constants. Dataset: *offers2.csv.*

Feature	Count	%Miss.	Card.	Min.	1stQRT	Mean	Median	3rdQRT	Max.	StDev.
ProductId	9886	0.0	307	-9.1726e+18	-4.35648e+18	6.104515668622242e+16	2.47757e+16	5.0561e+18	9.20715e+18	5.315584155400644e+18
ListingPrice	9886	0.0	4976	3.24	63.63	216.5	125.99	257.9	3194.3	256.6
SellerFeedbackRating	9886	0.0	25	0.0	91.0	88.97	95.0	96.0	100.0	21.6
SellerFeedbackCount	9886	0.0	123	0.0	338.0	6910.4	3293.0	8452.0	41420.0	10918.2
SellerId	9886	0.0	187	-9.13612e+18	-2.57228e+18	9.641003001213862e+17	1.16312e+18	5.47794e+18	9.17405e+18	4.919221779248733e+18
ShippingPrice	9886	0.0	1104	0.0	0.0	12.4	7.5	13.7	705.3	26.5
ShippingTime_minHours	9886	0.0	12	0.0	24.0	57.3	24.0	96.0	672.0	82.8
ShippingTime_maxHours	9886	0.0	11	0.0	48.0	88.9	48.0	120.0	1008.0	120.4

2 - CATEGORICAL FEATURES

Table 3. Descriptive statistics results for categorical features. Dataset: *amazon-offers-10k-samples-raw.csv.* Constant features (cardinality = 1) are in red.

Feature	Count	%Miss.	Card.	1stMode	1stModeFreq.	1stMode%	2ndMode	2ndModeFreq.	2ndMode%
IsWinner	10000	0	2	0	9451	94.51	1	549	5.5
TimeOfOfferChange	10000	0	548	2016-02-04T04:32:49.647Z	40	0.4	2016-02-04T04:38:26.096Z	20	0.2
ConditionNotes	5305	46.9	435	BRAND NEW…!!	758	14.3	"Brand new items…"	515	9.7
IsFeaturedMerchant	10000	0	2	1	8166	81.7	0	1834	18.3
IsFulfilledByAmazon	10000	0	2	0	9632	96.3	1	368	3.7
ListingCurrency	10000	0	1	CAD	10000	100			
ShippingCurrency	10000	0	1	CAD	10000	100			
ShippingTime_availtype	10000	0	1	NOW	10000	100			
ShipsFromCountry	6273	37.2	13	CA	3668	58.5	US	2213	35.2
ShipsFromState	5880	41.2	24	ON	2220	37.8	NY	1176	20
SubCondition	10000	0	1	new	10000	100			

Table 4. Descriptive statistics results for categorical features. Revised table after removal of duplicates, constants and features with high missing values %. Dataset: *offers2.csv.*

Feature	Count	%Miss.	Card.	1stMode	1stModeFreq.	1stMode%	2ndMode	2ndModeFreq.	2ndMode%
IsWinner	9886	0.0	2	0	9339	94.47	1	547	5.5
TimeOfOfferChange	9886	0.0	548	2016-02-04T04:38:26.096Z	20	0.2	2016-02-04T04:33:03.764Z	20	0.2
IsFeaturedMerchant	9886	0.0	2	1	8090	81.8	0	1796	18.2
IsFulfilledByAmazon	9886	0.0	2	0	9519	96.3	1	367	3.7

3 – DATA VISUALIZATION

Fig 1. Histograms for continuous features.

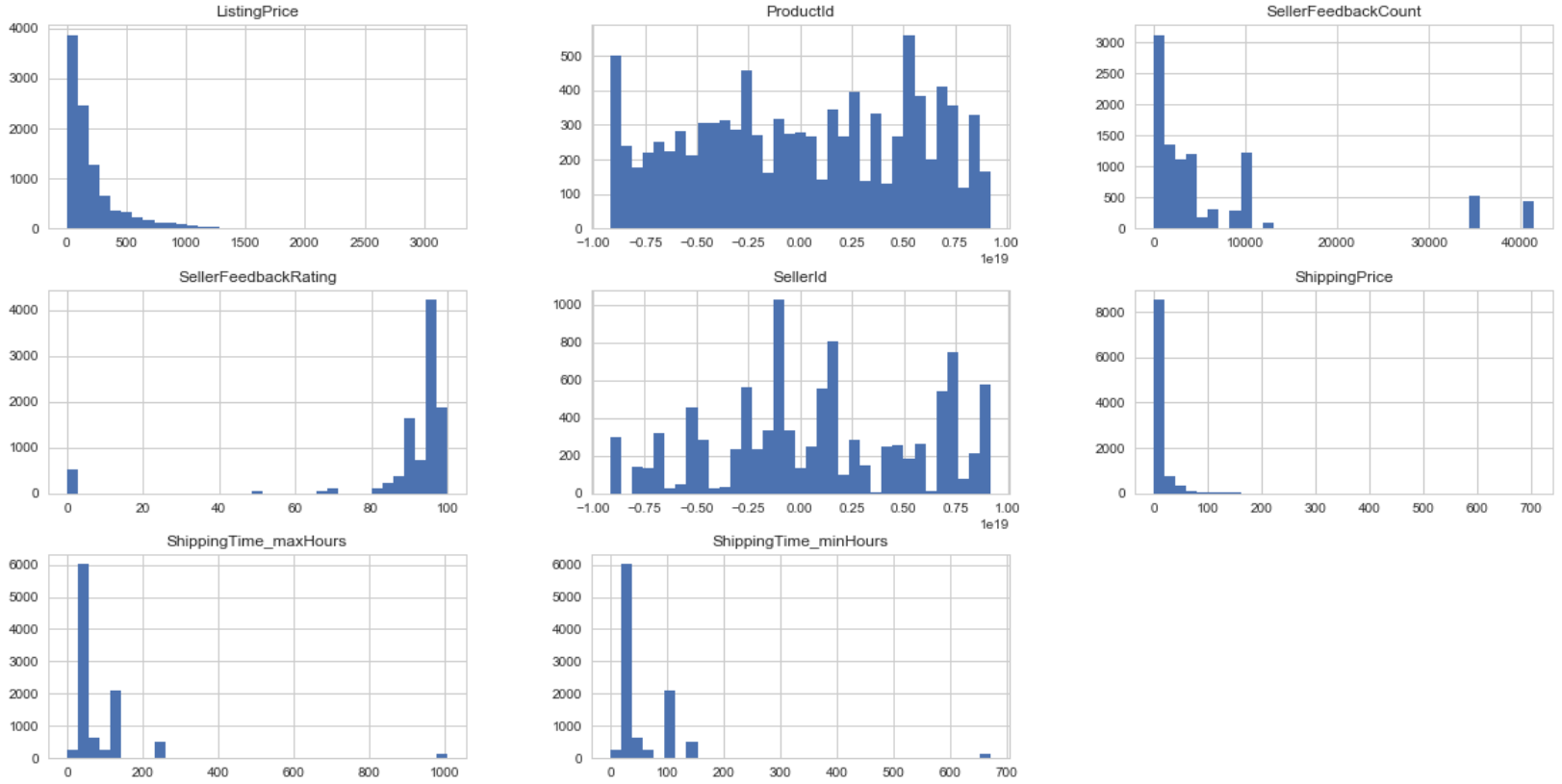

Fig 2. Box plots for continuous features.

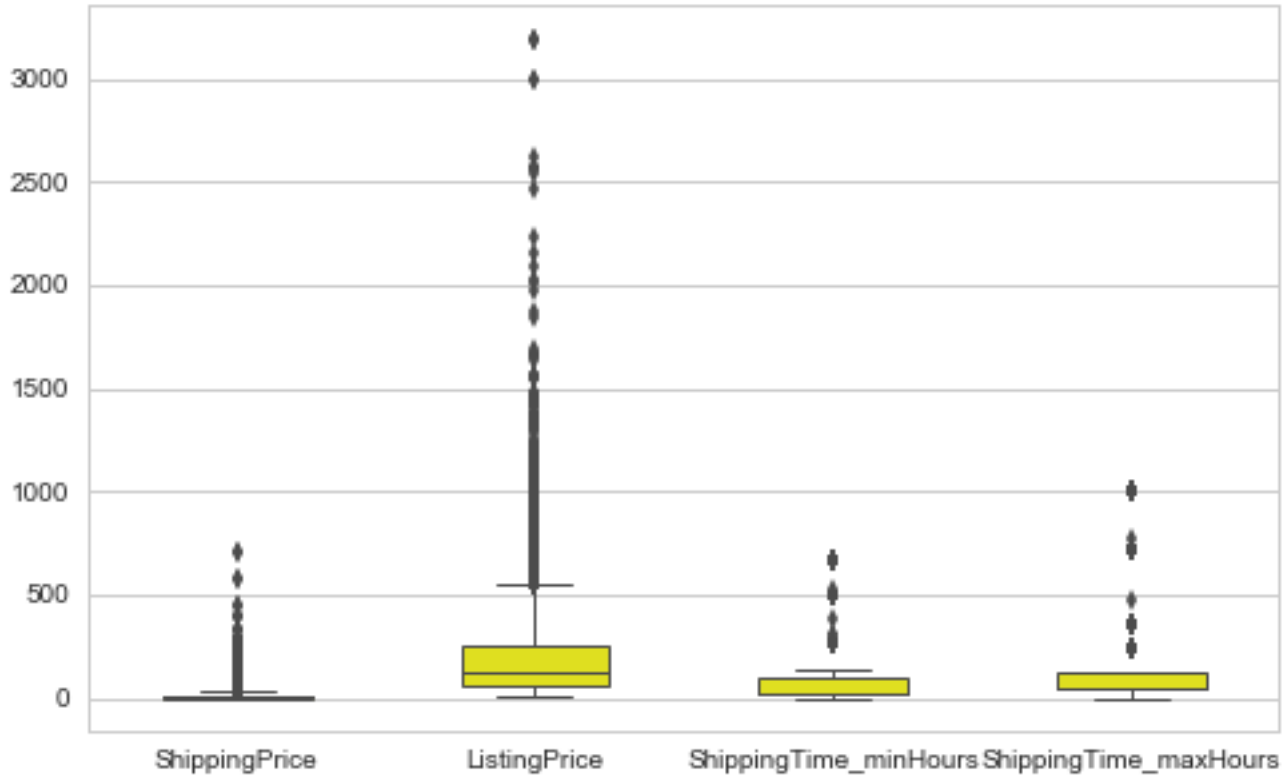

Fig 3. Box plots for continuous feature ‘ShippingPrice’ in more details, and box plots for the other features.

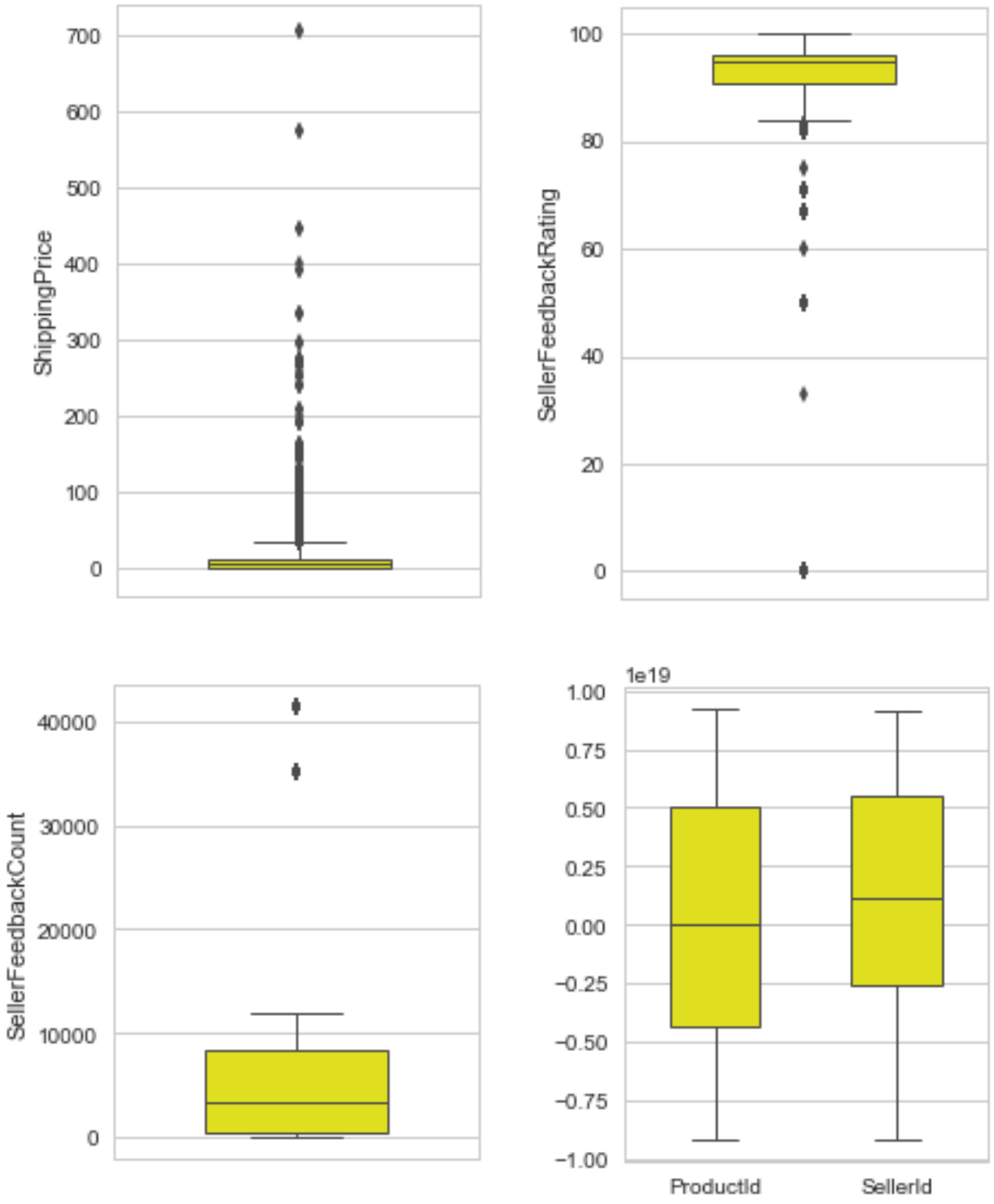

Fig 4. Bar plot for categorical feature 'IsWinner'.

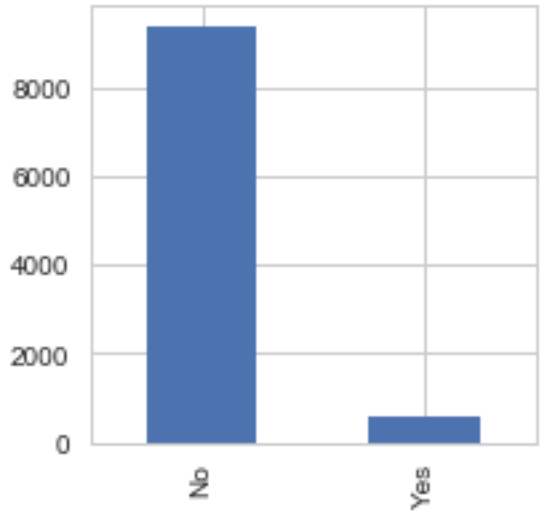

Fig 5. Bar plot for categorical feature 'IsFeaturedMerchant'.

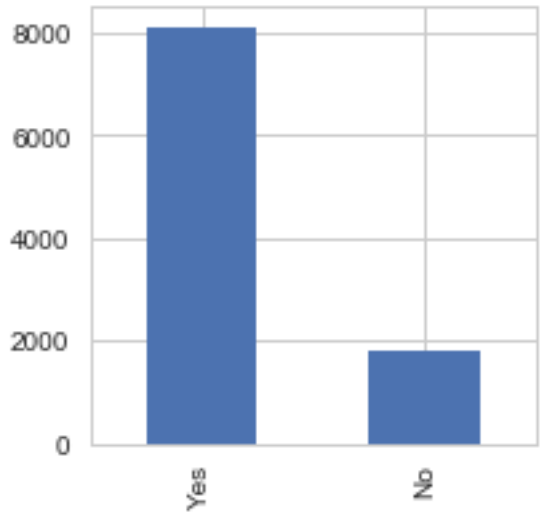

Fig 6. Bar plot for categorical feature 'IsFulfilledByAmazon'.

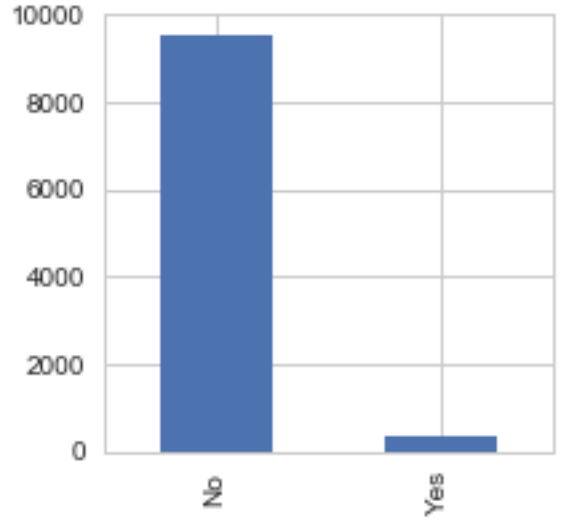

Fig 7. Scatter plot matrix for continuous features based on the target feature (Green dots for winner sellers, blue dots for non-winner sellers).

4 - CONSIDERATIONS

FROM HISTOGRAMS

- The features have different types of distributions.
- Exponential for ListingPrice and ShippingPrice.
- Skewed to the right for ShippingTime_minHours and ShippingTime_maxHours
- Skewed to the left (but almost exponential) for the SellerFeedbackRating.
- All these features could potentially have outliers, or errors, or misinterpreted missing values.
- SellerFeedbackCount could be bimodal, but not clear.
- SellerId and ProductId are both unimodal.

- SellerFeedbackRating presents a series of values at or around zero, isolated from the rest. It could indicate missing values? Or just relatively new sellers. Further inspection of the data is required. It is not clear why there is such a high difference between the data around zero and the rest of the data in the distribution, with mean at/around 90. Maybe the zero was used to indicate a missing value? If so, we could drop these values, or give them the value of the mean of the distribution, or drop altogether the feature if % of zero is too high (but from histogram seem only a small portion of the data).

- These considerations can be added to the Quality Data Plan for further evaluation later.

FROM BOX PLOTS

- We can see looking at the boxplots that many features present outliers, and we should deal with them. We take note of these in the Data Quality Plan (see file *DataQualityPlan.pdf*).

- As expected, the feature 'SellerFeedbackRating' present outliers near or at zero. We should further investigate how many of these outliers are present, then decide if to delete them, or assign to them a specific value (maybe at the lower range of the distribution, instead of using the mean). This feature should be treated with caution as it could be important in predicting the Winner sellers.

- Similar considerations should be made for the rest of the features where outliers are present, and register these on the Data Quality Plan, indicating a possible solution.

FROM BAR PLOTS

- The bar plots clearly define the two subsets in the various binary features. There are very few winner sellers in the data set, many featured merchants, and very little of the offers are fulfilled directly by Amazon.

FROM SCATTER PLOT MATRIX

- No clear correlations are seen between the descriptive features, but the division of data into winner sellers (green dots) and non-winner sellers (blue) gives some clear conclusions.
- Green dots concentrate in certain areas compared to blue dots. For instance, we can see that winner sellers have usually lower listing and shipping prices, but also higher ratings and lower minimum shipping time. No much differences were visible in rating counts.

- A good portion of values for seller ratings is at zero, but as we have seen, these are all outliers and will be dealt with in the Data Preparation Section.

Giuseppe Benanti – 04/03/2017