Giant Pumpkins EDA

Exploratory Data Analysis of the Giant Pumpkins data set

Summary of the data set

The data set of this project is from BigPumpkins.com. These statistics are from the Great Pumpkin Commonwealth's (GPC) Weighoff Results. It is sourced from the TidyTuesday Data project and the data set can be download here.

Each row of the data set represents a GPC weighoff result. Each row includes the id (year-type), place/ranking, grower name, city, state, country, gpc site and variety of the giant pumpkin. It also contains genetic info such as the seed mother and pollinator father. Measurements taken include the weight in lbs and ott in inches (Over the top measurement to estimate weight).

In the raw data, there are rows of 'seperator' inserted after the records of the same 'id'. We have removed these separator records and saved the processed data in the `processed_pumpkins.csv`. A script file for this data processing can be found in ????.

There are in total 28,011 observations and 14 features. Some null values found in the city, seed mother, pollinator father, ott, estimated weight, pct_chart and variety features.

Partition the data set into Training and Test sets

We will split the data with 70% training data and 30% test data. After splitting, the number of observations in the training set and test set are 19,607 and 8,404 respectively.

Exploratory Data Analysis on the Training set

We have plotted distribution of the target 'weight (Lbs)' and some features in the training set to explore if the features will be useful to predict the weight of the giant pumpkins.

The plot shows most of the observations are from the United States. The distribution of the GPC sites, city and state/province are more evenly distributed. We consider these columns are all good features to be used.

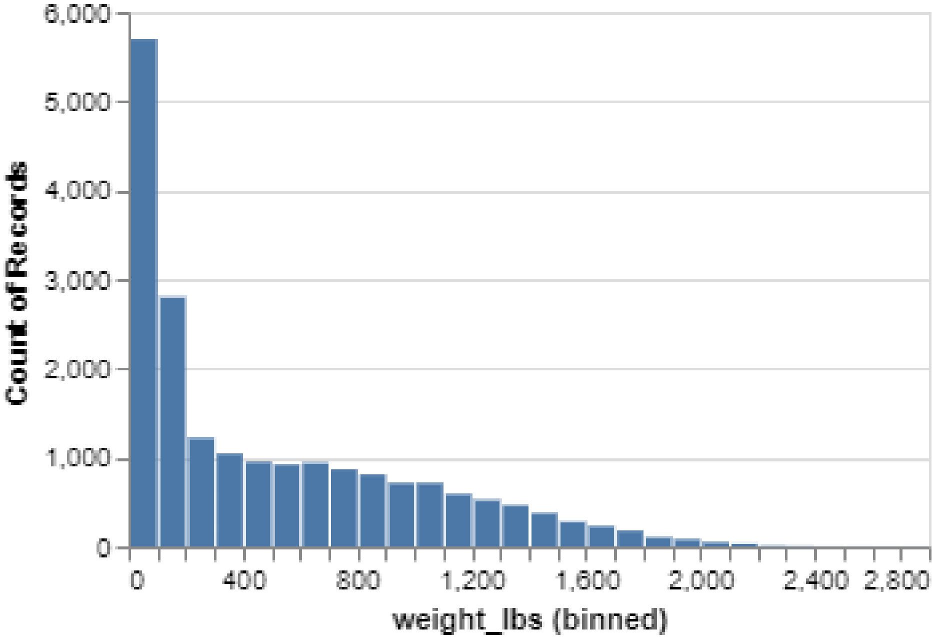


Figure 1: DIstribution of Weight (Lbs)

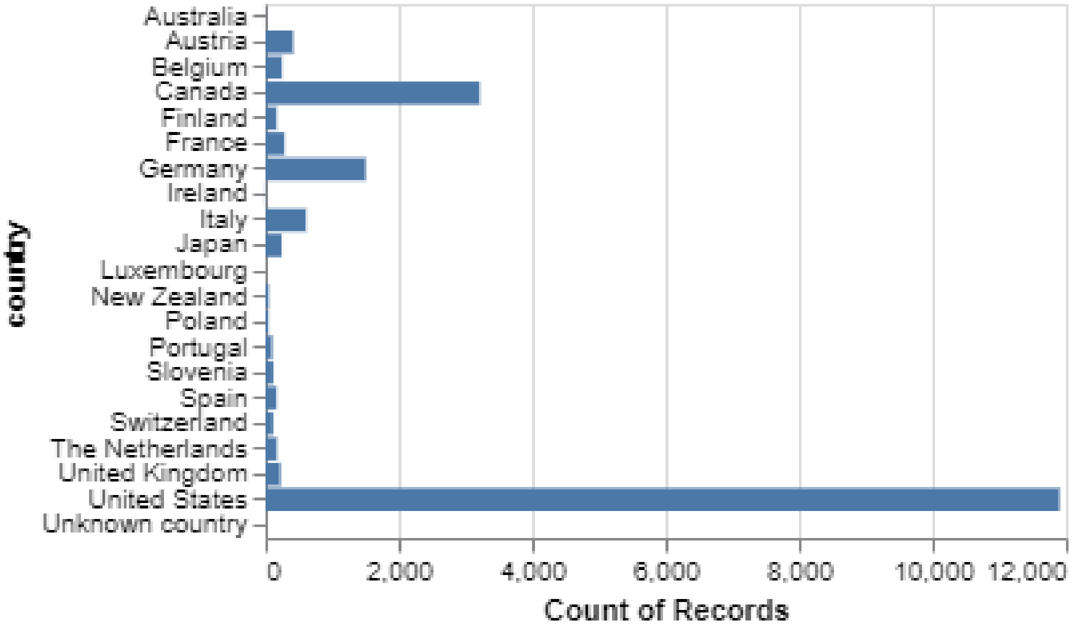


Figure 2: Distribution of Country

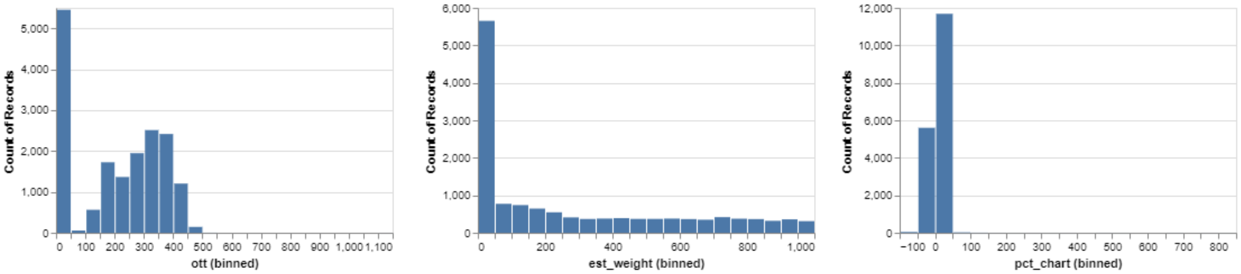
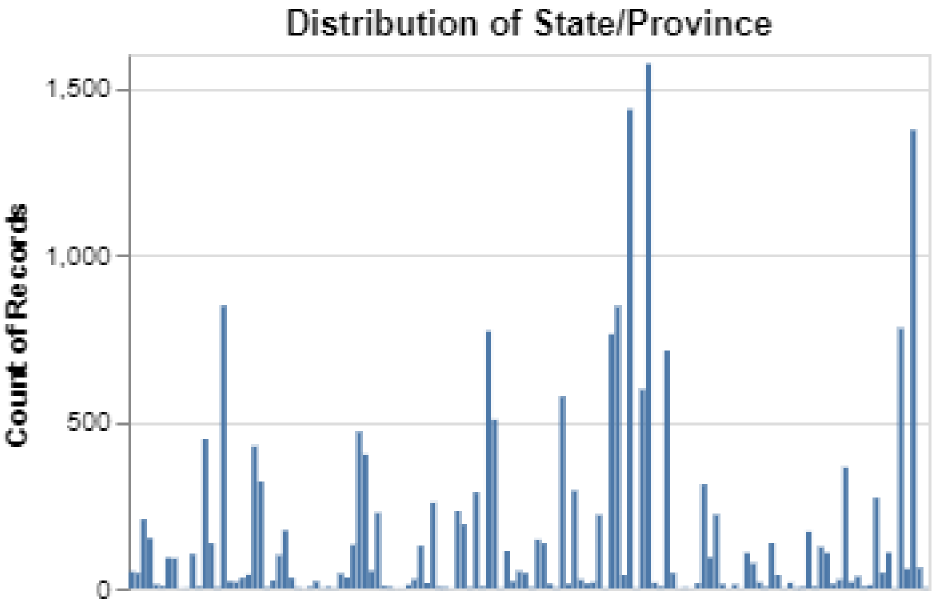
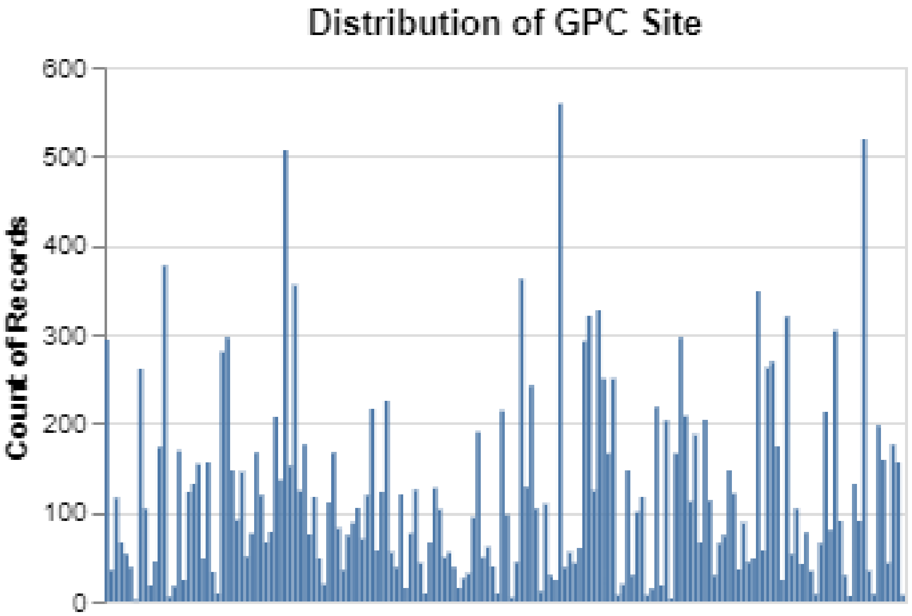
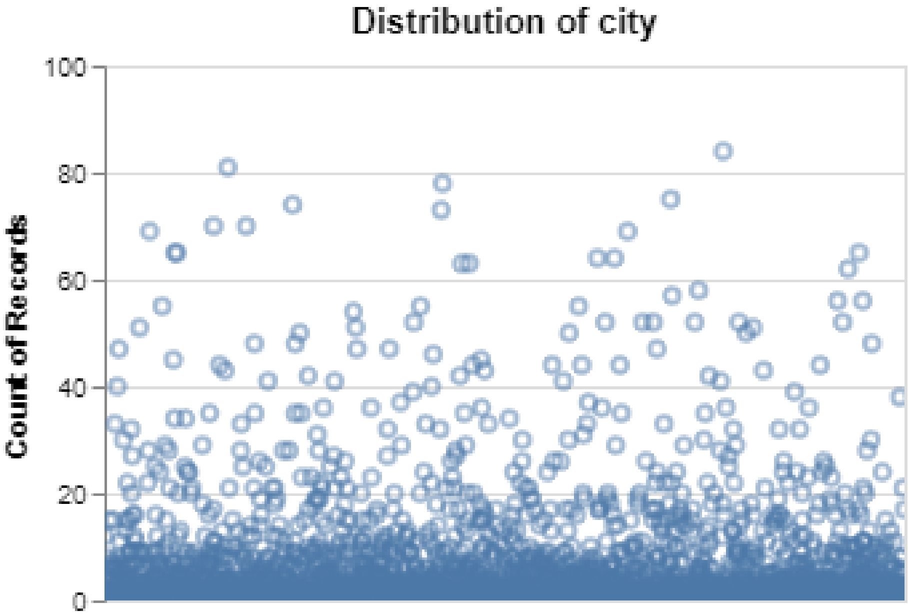


Figure 3: Distribution of ott, est_weight and pct_chart

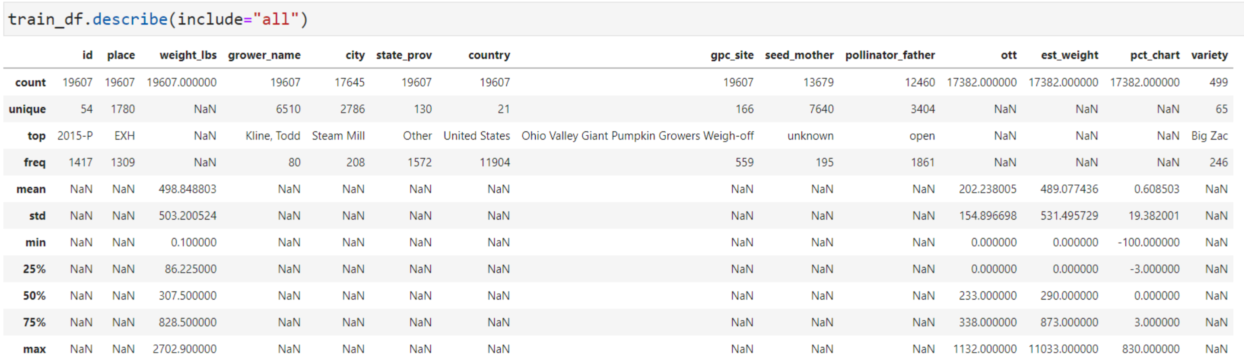

Distribution of State/Province


Distribution of GPC Site



From the describe summary of the training set (see below), it is noticed that the grower name, seed mother and pollinator father are free-text columns which we will use CountVectorizer to extract the bag of words for training. We think this genetic information might be useful for the prediction of the weight.

The number of non-null values in the variety column is very low. We will drop this column for training as the information may not be useful when there are so many null values.



```
train_df.describe(include="all")
```

	id	place	weight_lbs	grower_name	city	state_prov	country	gpc_site	seed_mother	pollinator_father	ott	est_weight	pct_chart	variety
count	19607	19607	19607.000000	19607	17645	19607	19607	19607	13679	12460	17382.000000	17382.000000	17382.000000	499
unique	54	1780	NaN	6510	2786	130	21	166	7640	3404	NaN	NaN	NaN	65
top	2015-P	EXH	NaN	Kline, Todd	Steam Mill	Other	United States	Ohio Valley Giant Pumpkin Growers Weigh-off	unknown	open	NaN	NaN	NaN	Big Zac
freq	1417	1309	NaN	80	208	1572	11904	559	195	1861	NaN	NaN	NaN	246
mean	NaN	NaN	498.848803	NaN	NaN	NaN	NaN	NaN	NaN	NaN	202.238005	489.077436	0.608503	NaN
std	NaN	NaN	503.200524	NaN	NaN	NaN	NaN	NaN	NaN	NaN	154.896698	531.495729	19.382001	NaN
min	NaN	NaN	0.100000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000	-100.000000	NaN
25%	NaN	NaN	86.225000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.000000	0.000000	-3.000000	NaN
50%	NaN	NaN	307.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	233.000000	290.000000	0.000000	NaN
75%	NaN	NaN	828.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	338.000000	873.000000	3.000000	NaN
max	NaN	NaN	2702.900000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1132.000000	11033.000000	830.000000	NaN

Figure 4: Output of Descriptive Summary of the Training Set from Jupyter Notebook