Wasif Mahmood Tahir

19100192

Assignment 1

Q1)

a) Missing values filled with average value.

Statistics:

Name	Type	Missing	Statistics				
				Min	Max	Average	Deviation
lowest_price	Real	0	Open chart	0	60	0.270	1.252
highest_price	Real	0	Open chart	0	2540	7.962	37.517

Filter (2 / 2 attributes): Search for Attributes

Missing Statistics:

Lowest price Variance = $(stdev)^2$ = 1.567

Highest price Variance = $(stdev)^2$ = 1407.525

b) Number of rows with invalid year = 67

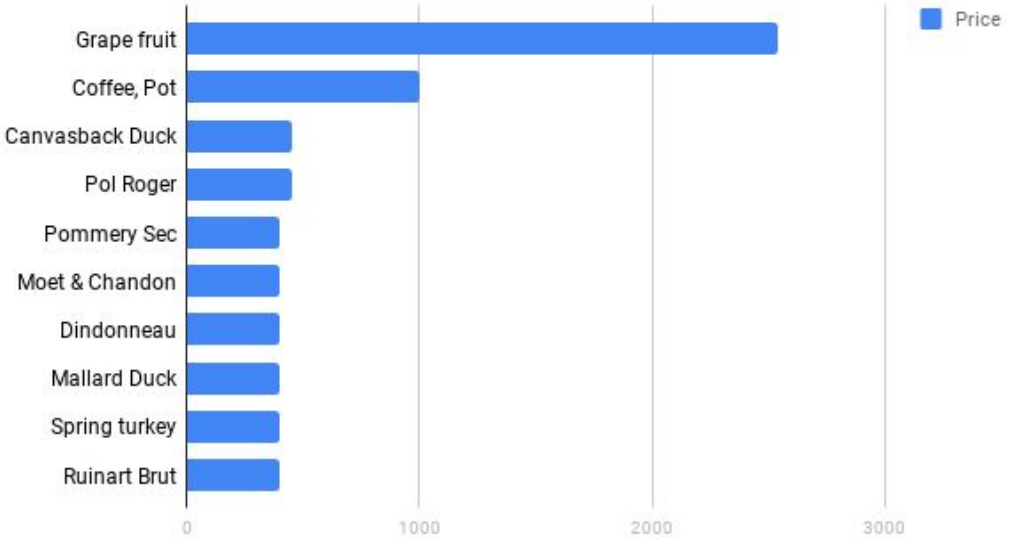
d) After part (a) the number of rows remain 10000

After part (b) the number of rows remain 10000

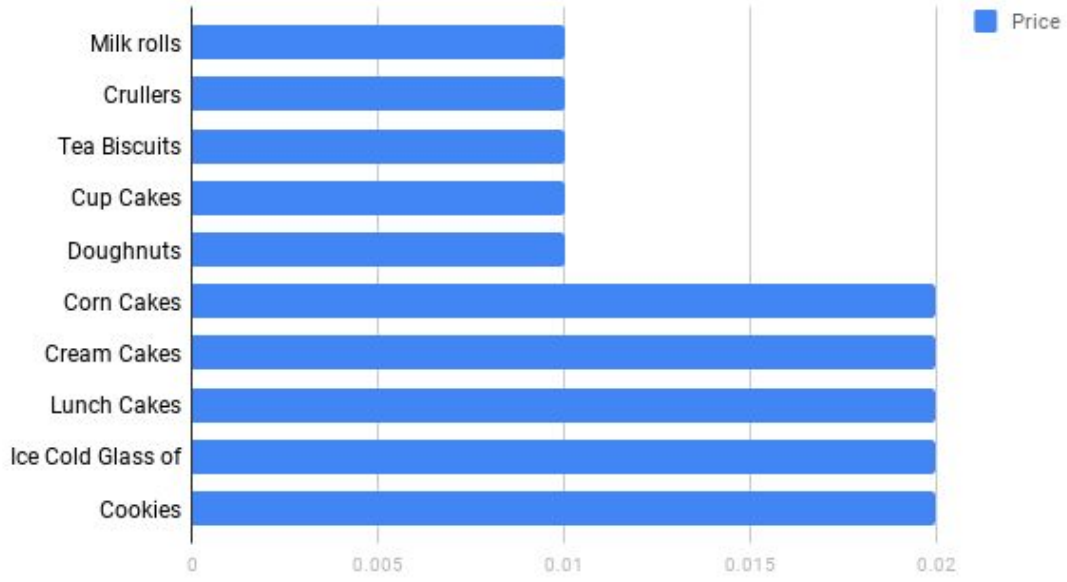
e) , f)

Top 10 Dishes	10 Most Expensive Dishes	10 Most Cheap Dishes
Coffee	Grape fruit	Milk rolls
Tea	Coffee, Pot	Crullers
Celery	Canvasback Duck	Tea Biscuits
Olives	Pol Roger	Cup Cakes
Radishes	Pommery Sec	Doughnuts
Mashed Potato	Moet & Chandon	Corn Cakes
Milk	Dindonneau	Cream Cakes
Boiled Potato	Mallard Duck	Lunch Cakes
Fruit	Spring turkey	Ice Cold Glass of Milk
Chicken Salad	Ruinart Brut	Cookies

Most Expensive Dishes



Cheapest Dishes



g) Top 5 dishes that appeared for the longest or shortest period of time in the menu:

Longest Appearing Dishes	Shortest Appearing Dishes
Ham	Cream of new asparagus, croutons
Macaroons	Striped bass saute, meuniere
Fried Egg Plant	Brook trout, mountain style
Radishes	Cerealine with Milk
Fruit	Wheat Vitos

h) The best representation depends on what type of analysis we are doing. For tasks like comparing prices a bar graph would suit best, while for classifying values into groups a scatter plot may make more sense.

Q2)

a) Number of missing values for the first 10 attributes:

Attribute	No. of Missing Values
#1	0
#2	1174
#3	1177
#4	0
#5	0
#6	0
#7	0
#8	0
#9	0

#10	0

Top 5 attributes with the highest number of missing values:

Attribute	No. of Missing Values
102	1675
103	1675
104	1675
105	1675
106	1675

b) Missing values filled with average value.

c) Dataset normalized.

Q3)

a) The following attributes from the correlation matrix were had a significant correlation value with the class label [Result of Treatment].

	Sex	Age	Time	Number_of_Warts	Type	Area
Value	-0.086203	-0.542780	-0.654147	0.078273	-0.485030	-0.188886

The following attributes from the correlation matrix had a significant correlation value with other attributes (other than the class label).
(Considering the values of the data, a correlation value of 0.3 will be considered as a significant correlation)

Attribute A	Attribute B	Value
Type	Age	0.415536

Area	Type	0.354398

b) Chi Square plot of Age against Number_of_Warts:

Number_of_Warts	1	2	3	4	5	6	7	8	9	10	11	12
age												
15	1	3	1	0	1	0	1	0	0	1	2	2
16	1	0	2	0	0	0	0	0	0	0	0	0
17	0	2	1	1	0	0	0	0	0	1	0	1
18	2	0	0	0	0	0	0	1	0	0	0	0
19	0	0	0	0	0	1	0	1	1	0	0	0
20	0	1	2	0	0	1	0	0	0	0	1	1
21	0	0	0	0	2	0	1	0	0	0	0	0
22	0	2	0	0	0	0	0	0	1	0	0	0
23	0	0	0	0	1	0	1	0	0	0	0	1
24	0	0	1	0	0	0	0	1	0	1	0	0
27	0	1	0	0	1	0	1	0	0	0	0	0
28	0	0	1	0	0	0	0	0	1	0	1	0
29	0	0	0	0	2	1	0	0	0	0	0	0
30	0	1	0	0	0	0	0	1	0	1	0	0
32	0	0	0	1	0	0	1	0	0	0	0	1
34	3	0	3	0	0	0	0	0	0	0	0	0
35	0	1	0	0	2	1	0	1	1	0	0	0
36	0	1	0	1	0	1	0	0	0	0	0	0
40	1	0	0	0	0	1	0	0	1	0	0	0
41	0	2	0	0	1	0	0	0	0	0	0	0
50	1	0	0	1	0	0	0	0	0	0	1	0
59	0	0	1	0	0	0	0	0	0	0	0	0
63	0	0	1	0	0	0	0	0	0	0	0	0
67	1	0	0	0	0	0	1	0	0	0	1	1

Chi-square Statistic: 258.976648352
p_value: 0.384678789512

c) There are varying correlations between the attributes and the class label. Time, Age and Type have the highest magnitude of correlation which means they have the most weightage in determining the Result_of_Treatment.
Other attributes with any considerable correlation are {Area, Type} = 0.35 and {Type, Age} = 0.41
The Chi Square Statistic value is large which means the observed and expected differ greatly and the null hypothesis of independence is rejected. The high p-value also corroborates the rejection of the null hypothesis.