Modern   
Data science

Assignment 2

Saurabh Dharmadhikari

Saurabh.dharma01@gmail.com

Question – 1

Table of Content:

|  |  |
| --- | --- |
| **Question 1.1**  Find the missing values: • Write the function missing\_values\_table and use the dataframe as the input. The function should return the information of missing values by column (only for columns which have missing values and the returned value should be the count of rows has missing values); • For columns which have missing values, could you impute the missing values with the mean value of the particular columns? (if you think it could not be done with mean value, write down the reason in comments and report rather than code) | 2 |
| **Question 1.2**  Find the price information from the data: • Write code to print the median price of the items in the data; • What is the 90th percentile value on the price; • Draw the histogram chart for the price of the items in the data with 50 bins | 2 |
| **Question 1.3**  Exploring the shipping information from the data: • Write code to find out the percentage of the items that are paid by the buyers. • Draw (two) histogram graphs in one plot on the price for seller pays shipping and buyer pays shipping (50 bins). • When buying the items online, do you need to pay higher price if seller pays for theshipping? Write the code to find out (Compare the median price of items paid by buyers and items paid by sellers, and explain the result in the comment and report). (Optional: You could use the subplot from EDA) | 2 |
| **Question 1.4**  You are required to find out the item condition information from the data. Lower the number (value), the better condition of the item. • Write the code to find out (print) the count of the rows on each number (value) in column item\_condition\_id. • Draw the boxplot graphs (one plot) on the price for each item condition value, and find out out whether the better condition of the item could have higher median price (draw the plot and answer this question in the comment and report). | 3 |
| **Question 1.5**  Conduct the category analysis and find out the relevant information: • Write the code to find out (print) how many unique categories you could find from column category\_name. • For the items with worst condition only (highest value from item\_condition\_id), write code to (print) find out the top 3 categories (now you probably understand the findings you had in Question 1.4).  **Question 1.6** The categories in column category\_name have 3 parts. The three parts (main\_cat,subcat\_1 and subcat\_2) are concatenated with ’/’ character sequentially in the data now. • Write the function (must be function) to split the text content (string value in each row) in column category\_name by ’/’ character. you need to handle the exception in the function for those has missing values (NaN). For missing values (NaN), the results from splitting should be ”Category Unknown”, ”Category Unknown”, ”Category Unknown”. • Use the above function you wrote to create three new columns main\_cat,subcat\_1 and subcat\_2 with corresponding values from the result of splitting. Print out the dataframe to show the top 5 rows for three new columns main\_cat,subcat\_1 and subcat\_2. |  |
| **Question 1.7** After splitting the category for column category\_name, we now have the three main details regarding to the category information. However, we need to clean the text in each of the new three columns in lowercase. • Write code (or function) to change the text (value in each row) from the new three columns to lowercase. • Draw the bar chart to find out the top 5 most popular main categories (in column main\_cat) in the data (only showing the top 5). • Write code (or function) to (print) find out how many unique main categories (in col umn main\_cat), unique first sub-categories (in column subcat\_1) and unique second sub-categories (in column subcat\_2) respectively. |  |
| **Question 1.8** Exploring the price and categories. • Write code to (print) find out the median price for all the categories in new column main\_cat. • Draw the bar chart to find out the top 10 most expensive first sub-categories (in column subcat\_1) in the data. • Draw the bar chart to find out the top 10 cheapest second sub-categories (in column subcat\_2) in the data. |  |
| **Question 1.9** Exploring the price and brand. • Write code to (print) find out the median price for all the brands (fill NaN with ’brand unavailable’). • Draw the bar chart to find out the top 10 most popular brands in the data. |  |
| **Question 1.10** Item Description Analysis. • Could you draw the wordcloud chart by using the column clean\_description. • Divide the data with quantiles of the price (using qcut from pandas to obtain the first/sec ond/third/fourth quantile). • Draw the wordcould by using the column clean\_description on each quantile of price data. |  |