Kyle Calabria Springboard Data Science

12/19/2017 Capstone One

Capstone One Milestone Report

**Define the problem**

The problem I selected was a predefined practice problem from a datahack hosted by Analytics Vidhya called Black Friday. The goal of the problem is to predict customers next month purchase amount for a collection of popular products based on their previous months purchase amounts for the same collection of products.

**Identify your client**

The client is an undisclosed online retail company “ABC Private Limited”. Although the problem does not directly state their motivation for predicting their customers upcoming purchasing behavior, it is likely that they are motivated to either properly stock their inventory or to improve their targeted advertisements, promotions, and upselling.

**Describe your data set, and how you cleaned/wrangled it**

My dataset contained twelve fields and 65,499 records. Of the twelve fields, eleven were independent variables (User ID, Proudct\_ D, Gender, Age, Occupation, Gender, Age, City Category, Stay In Current City by Years, Marital Status, Product Category One, Product Category Two, and Product Category Three) and one was dependent variable (Purchase). The dependent variable purchase, contained the currency amount purchased of the product by the user in the previous month.

The first step I took in cleaning the data was to remove or impute missing data. The only product categories with missing data were Product Categories two and three. Since this was categorical data, I assigned all NaN values to 0 to represent no category.

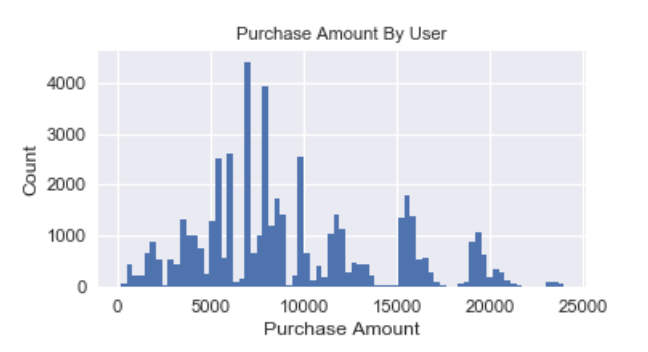
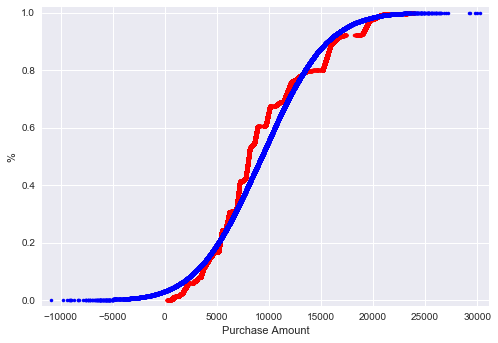
The second step I took in preparing my data was to hot one encode all of the categorical data. I used the OneHotEncoder, reshape, and fit\_transform functions from sklearn to encode the following categorical data fields: gender, age, occupation, city category, and stay in current city years. While hot one encoding the occupation data I had to interpret occupation 0 as either a category of occupation or as a missing data value. From examining the data, the number of people with occupation 0 fell within the range of people with other types of occupations so I concluded that occupation 0 was a category of occupation and did not remove records with occupation 0. Hot one encoding product category data was slightly more complicated that the previously mentioned variables, because I had to aggregate three fields of data before hot one encoding. To complete this, I turned each product category into a matrix with columns representing product categories (20) and rows representing records. I then added the three matrices together to capture all product category data in a single matrix. The final step was appending all the hot one encoded data to the initial data frame and clean up column titles.

The third step was splitting my data into a training and test set. I assigned the first 75% of data into the training data set and the remaining 25% of the data into the testing data set. I then saved the two datasets into separate csv files and began the eda phase of my project.

**List other potential data sets you could use**

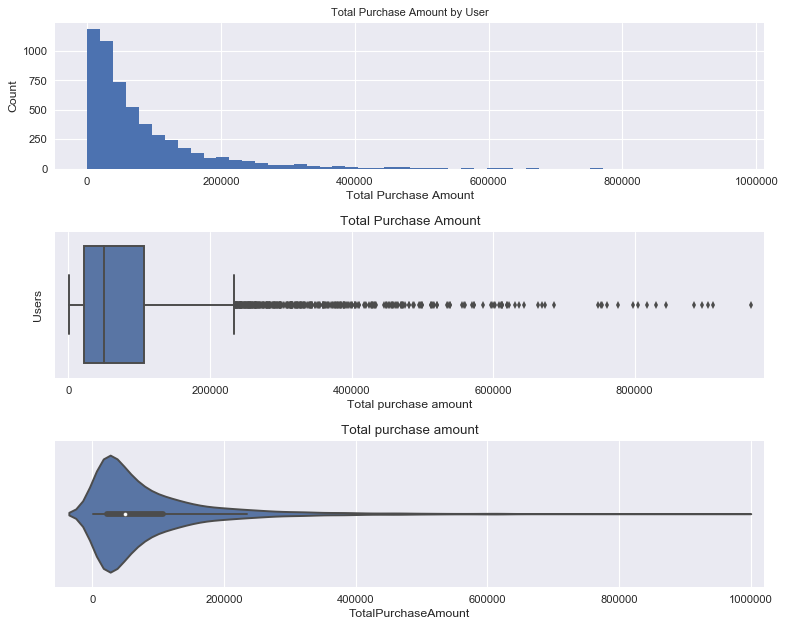
Given that this project was a predefined problem from a hackathon with an associated dataset it is difficult to list other possible data sets. Selecting a predefined problem with a known and high quality data set made practicing introductory data science skills easier. Having said that, this dataset was not ideal, because it did not contain many numeric data fields nor did it contain any time or data fields. This limited my ability to conduct quality linear / multilinear regressions or conduct any time series analysis. Other potential data fields that the problem creator could have included would be: date and time of purchase, geolocation or purchase, multiple months or even years worth of data, discounts used, items purchased in the same basket, and length of time the user has been registered on the site.

**Explain your initial findings**

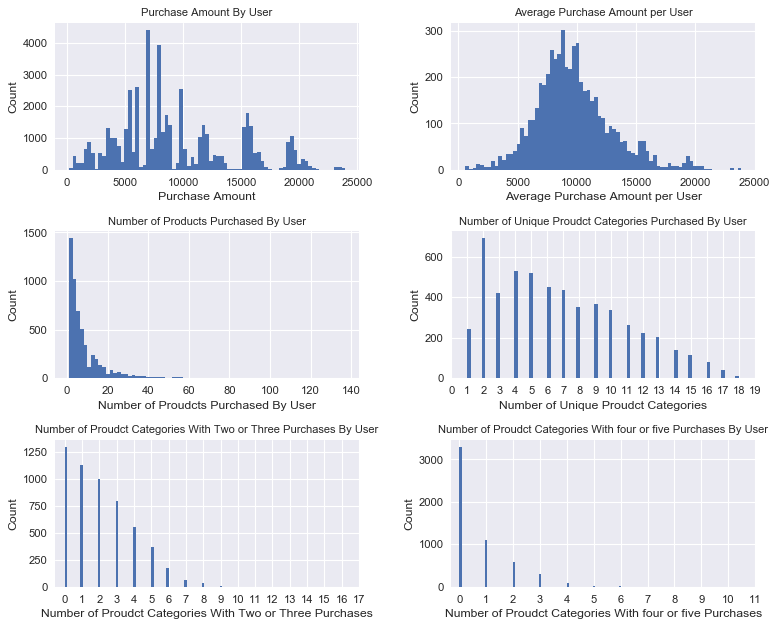
 

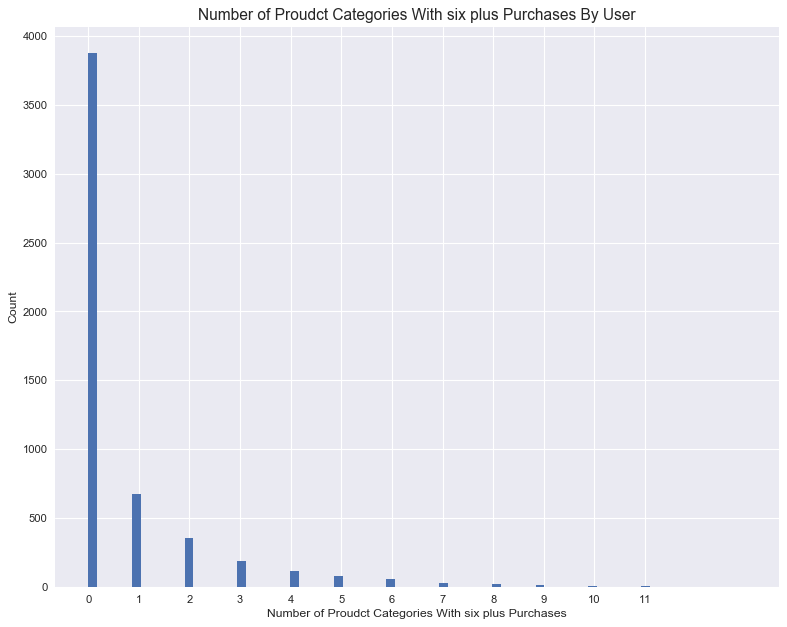
The training data set has 49,124 records with an average purchase amount of 9,286.17, a standard deviation of 4,954.87 and a median purchase amount of 8,048.0. The data set does not follow a normal distribution, with a p value of 0.00 on the normalcy test. The cumulative density function above on the left visualizes the data sets diversion from normalcy. From the density mass function above on the right the data set seems to be multi-modal and positively skewed. There appears to be clustering for purchase amount by user. This might suggest clusters of products that cost similar amounts or popular products i.e. macbook pros and Iphones.

Below portrays the data on a total purchase amount by user basis. The average total purchase amount per user was 84,304, the median was 50,030, and the standard deviation was 100,918. The data was highly positively skewed with a standard deviation greater than the mean indicating a significant number of outliers. This is also visualized well in the box and whiskers plot below.

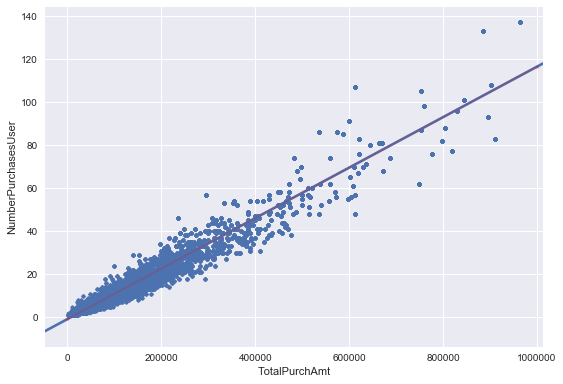
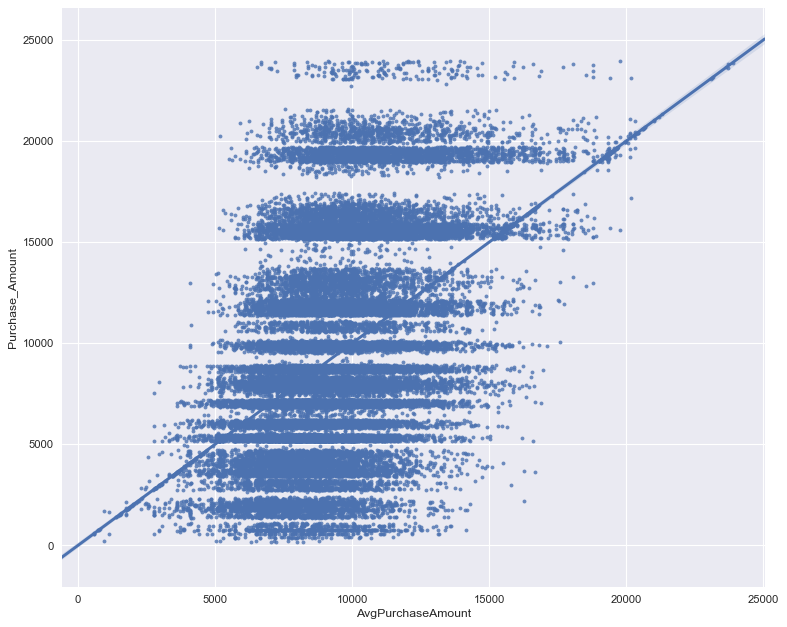


The number of products purchased by user seems to decline along an exponential curve, suggesting the large majority of user purchase between one and eight products. Two is the most common number of unique product category purchases per user (middle right plot) which seems unusual that one would not be higher. This is likely due to the fact that many categories have multiple product category tags, which slightly reduces the descriptive ability of this unique category field. The unique product categories purchased histogram generally follows a downward linear trend with a few anomalies being three categories and eight categories having fewer than expected. The number of categories with two or three purchases follows a steeper downward curve than unique categories, but still seems linear, whereas four or five purchase categories and six or more purchase categories seem to follow exponential downward curves.





This data set does not scatter plot well due to its lack of numeric data types. The most highly significant trend line appears in the plots of purchase amount and average purchase amount and in the plot of number of purchases and total purchases. There is a positive correlation between a user's purchase amount and their average purchase amount which makes sense as average purchase amount is a function of the user's aggregate purchases. There is also a very high positive correlation between the total purchase amount and the number of purchases. This suggests that the variance in total purchase amount amongst users is greatly in part due to the number of purchases as opposed to a higher average purchase amount. The pearson correlation for this plot was 0.966.



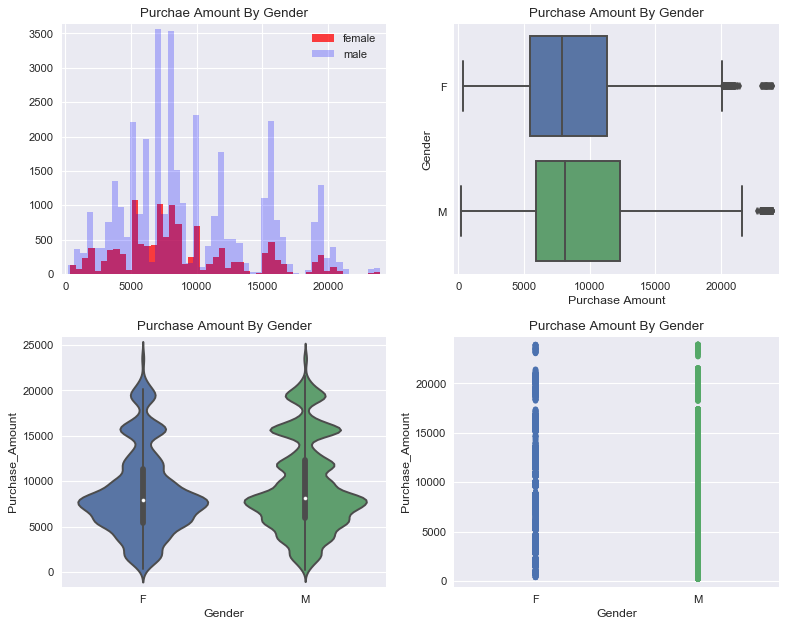
After taking statistically analyzing the data on a by user basis, I examined the data by grouping the data based on a variety of categorical features i.e. gender, age, occupation, and city category.

**Gender**

The average purchase amount for males was 9,464, the median was 8,099 and the standard deviation was 5,026. The average purchase amount for females was 8735, the median was 7913, and the standard deviation was 4683.

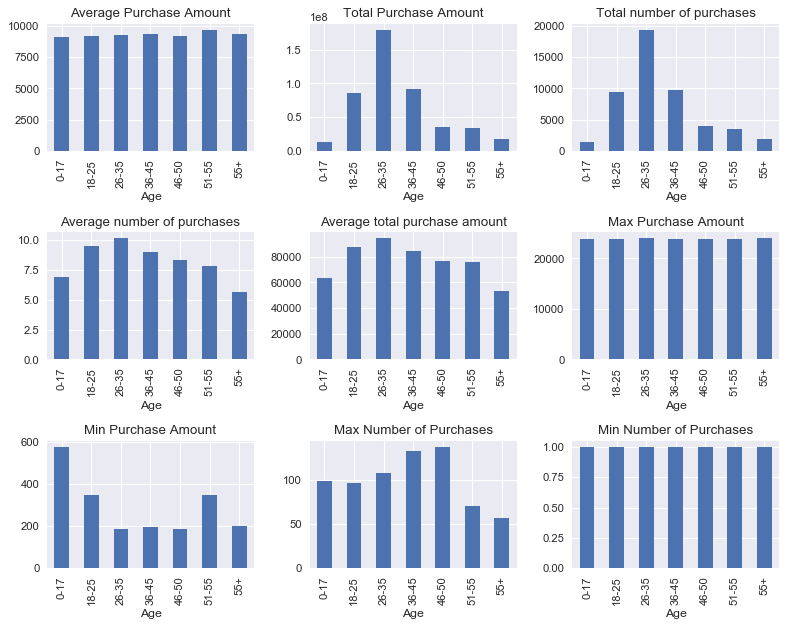
The average purchase amount, average number of purchases and average total purchase amount are marginally higher for males than females, but the total purchase amount and the total number of purchases is roughly 3 times larger by males than females suggesting that males are the primary purchasers on this particular online retailer. The max purchase amount is equal for the genders, but the minimum purchase amount is about 2 times smaller for men than for women. The maximum number of purchases is roughly 25% higher for men and the minimum number of purchases for both purchases is one.

The overlay of average purchase amount per gender shows similar mean purchase amounts for the genders, but roughly three times as many purchases for males than for females. From the violin plot it is clear that males tend to make more purchases at the 15k range.

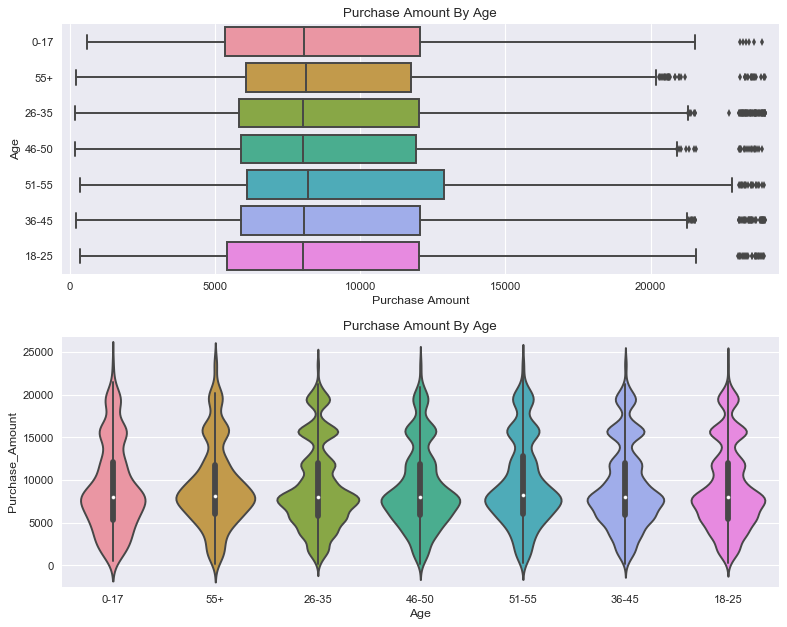


**Age**

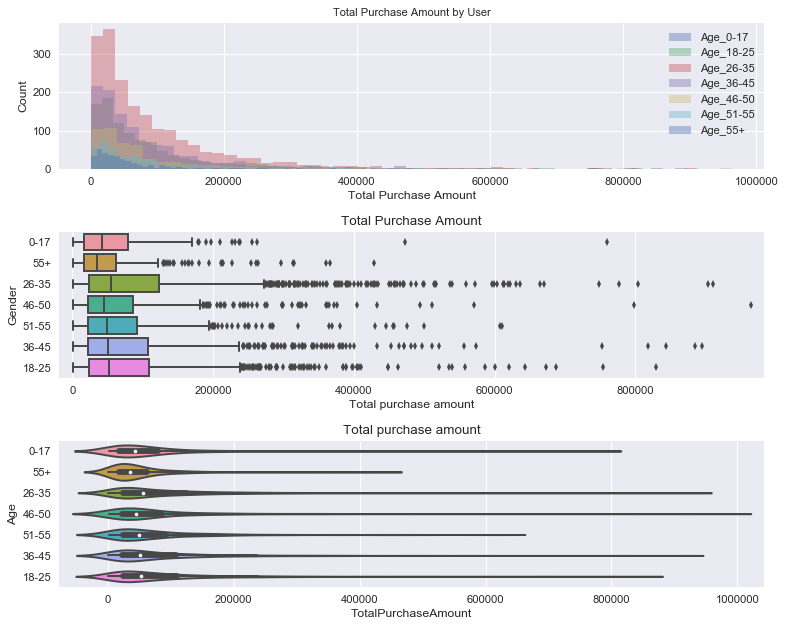
The average purchase amount is similar across age ranges, but the total purchase amount and number of purchases varies dramatically with the age range. Age range 26-35 far out purchases all other age ranges and 18-25 and 36-45 have similar total purchase amounts tying for second place. The remaining age ranges purchase far less. The average number of purchases and average total purchases have similar ordering of age ranges in terms of quantity but the quantities are much closer together.



Examining the violin plots of the purchase amount by age shows that ages 26-35 have a significantly greater proportion of purchases in the 15,000 and 20,000 price clusters than other ages. Ages 36-45 and 18-25 also have significant amounts of purchases in the 15,000 and 20,000 price cluster whereas ages 0-17 and ages 55+ have much fewer purchases in the upper price ranges.

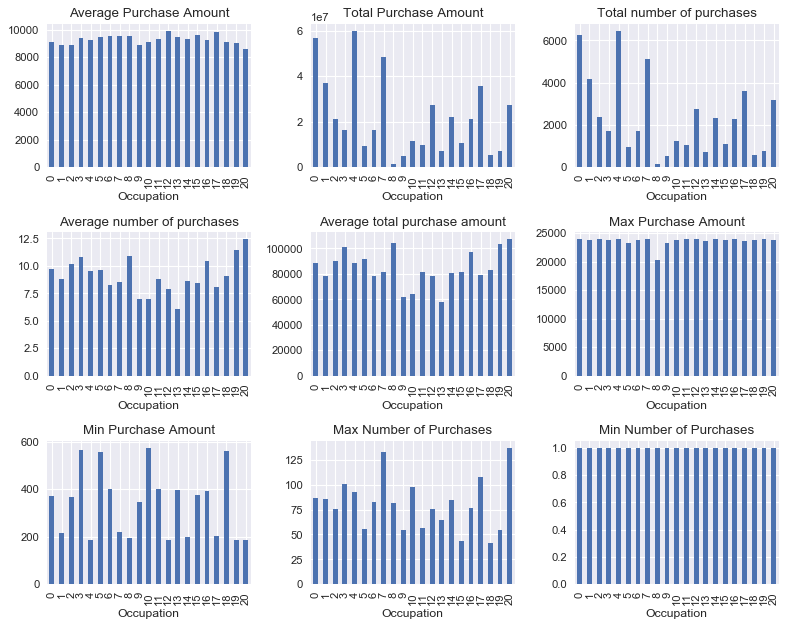


Examining the below plots of total purchase amount by age, it is clear that 55+ has by far the least total purchase amount on average. It is also interesting to see how many outliers that are specifically in the 26-35 age range but also in the 36-45, and 18-25 age ranges.



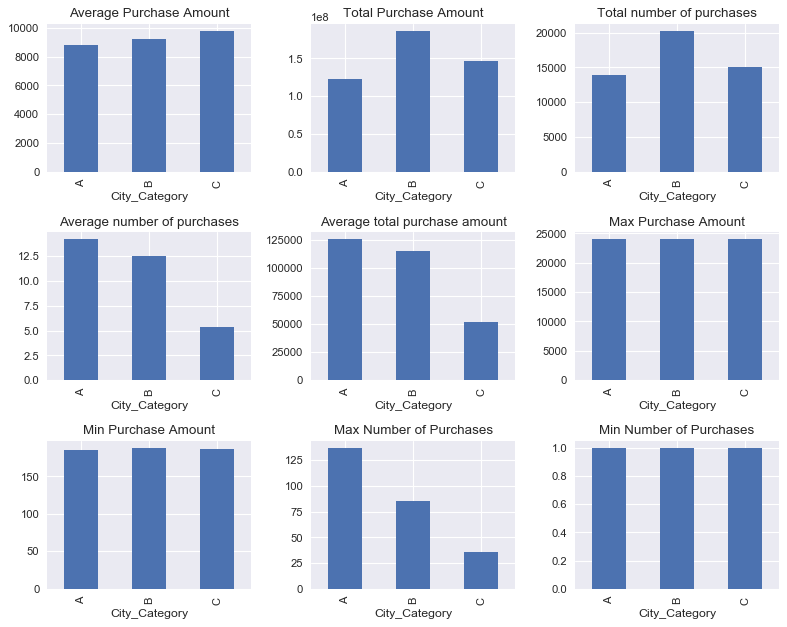
**Occupation**

Average purchase amount, maximum purchase amount and minimum purchase amount seem to be relatively similar across occupations. Total purchase amount and total number of purchases have a wide range of amounts and a near identical distribution. Occupations 0, 4 and 7 are the top three purchasing occupations and occupations 8, 18 and 19 are the bottom three purchasers with occupation 8 being significantly below even the second lowest purchaser. Average number of purchases and average total purchase amount follow a very different distribution suggesting that total purchase amount and total number of purchases reflect more on the number of people that fall within each occupation as opposed to the purchasing habits of people in each category. The Average number of purchases and average total purchase amount are more uniformly distributed and reflect more accurately the purchasing characteristics of an occupation. In both of these charts, occupation 8 has the second or third highest amount.

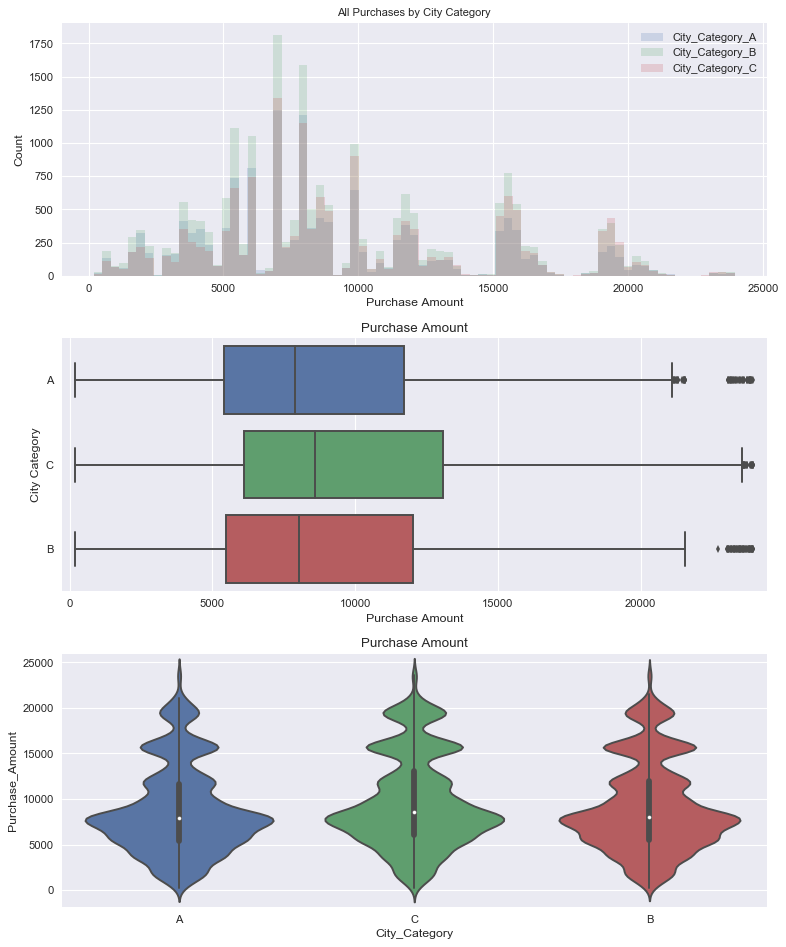


**City Category**

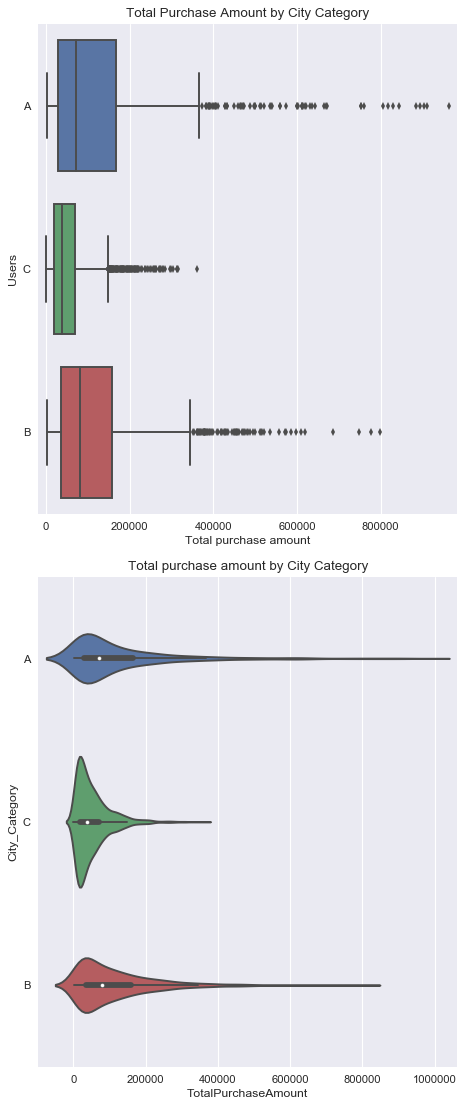
City Category A has the lowest average purchase amount, total purchase amount and total number of purchases, yet the highest average number of purchases and average total purchase amount. This would suggest a higher spending (more affluent), less populated community.



Examining the below violin plot of purchase amount by city, it seems that although city group A spends more per capita than city group B or C, city categories B and C spend more in the upper pricing clusters of ~15,000 and ~20,000. This would also suggest that group A spends a higher proportion of their overall spending near the median price of ~7,903.

****

Examining the below total purchase amount by city box and whisker it is evident that although C had expensive single item purchases, their total purchase average and interquartile are significantly below A and C. Also noteworthy is the amount of outliers in city group A in general as well as above 80,000 which is well over three interquartile ranges away from the average.

****

**Model Building**

Running data across multiple platforms

12 data frames vs 4 data frames vs 1: outliers, bin vs non--bin,

Feature importance

Differences in models

Feature trimming: RFR

Feature engineering: product average, high flag, etc.

Encoding of data

Types of models

Appendix