Project: Capstone Project 1: Milestone Report

Unit 8.5

Problem Statement

I am looking to understand customer spend amounts across various product categories. Based off customer spend data for a month, which product category should be showcased in a personalized offer to a customer that will drive a sale in the product category. A model to predict the purchase amount of the customer against the various products will help answer this question.

In today's retail landscape, consumers have an unprecedented amount of choices & convenience on where, when and how to shop. Loyalty programs have become a way for retailers to harness customer data and drive repeat trips. Retailers can lean into this data to help anticipate customer needs & spend to provide unparalleled service by anticipating needs of the customer through personalized offers on what matters most to the customer, providing options for a customer before having that need met by a competitor.

Along with predicting the purchase amount of customers against various products, I will also predict the likelihood a customer will purchase certain categories based on demographic information.

This will move beyond current customers and anticipating their needs, to marketing to potential customers that not have yet shopped at a company. This can help with targeted ads for companies to know where they should spend their marketing fund to attract new customers, by marketing personalized ads to drive traffic.

Data Wrangling

The dataset is 550,068 rows and 11 features and the target variable. All data features are categories except for the Purchase target variable.

The features Occupation, Product Category 1, Product Category 2 and Product Category 3 are all masked. A Product ID can have more than one Product Category designation. This is a bit challenging to decipher with the categories being masked. This can either mean that a '1' in Product Category 1 and Product Category 2 is the same category, for instance 'shoes'. Or if they were independently masked, even though the number '1' mask is the same, they are different categories.

I am going to make the assumption that these are sub categories of Product Category 1, which may be seen in a retail store. For instance, a retail store may have sub categories to help designate items within a category (ie. Accessories > Handbags > Wallet).

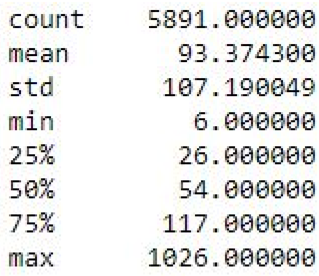
After importing the data, I took a few steps to inspect/clean my data. I checked the data types for the different columns using the .info() method. I found that all of the columns that had values, such as ‘Purchase’ were indeed int or float data types. I also checked for missing values, and found that the only missing values were in the Product Category columns, and that is to be expected, every customer did not purchase every type of product on each shopping trip due to this, I chose to not to action on these missing values.

I noticed that a few of the columns would be better as categorical data types because there are not that many different options. This can help by speeding up the groupby methods and utilize less RAM. I chose to import the columns: Gender, Age, Occupation & Marital status as category data types.

I checked that each of the columns were labeled without erroneous spaces
Most of the columns are categorical therefore there were no outliers. I checked the Purchase amount column - there seem to be a few outliers. I did not do adjust the outliers, this is useful data to investigate the correlation of the higher spend with other variables to help answer the overarching question of increase spend at this retail location.

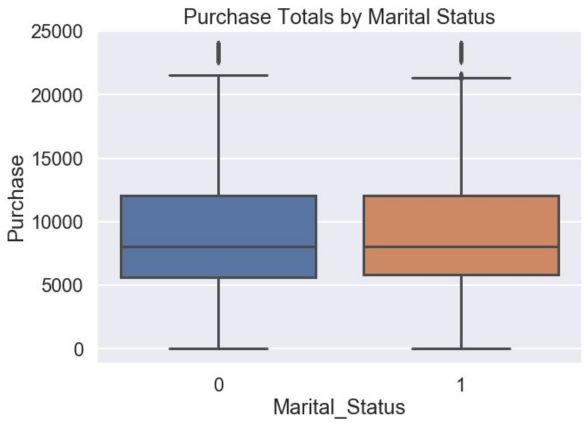
Inferential Statistics

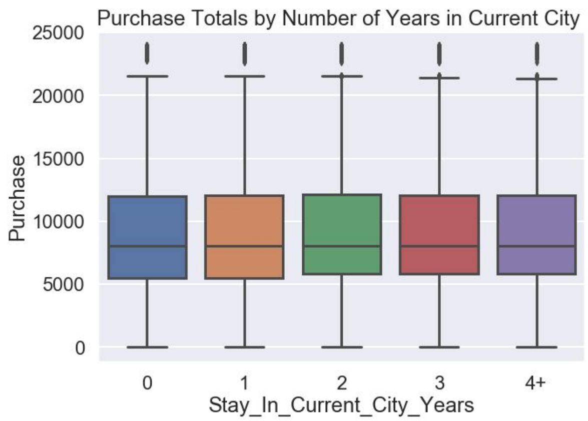
The average customer purchased 93 products over the time period of the data, with the top customer purchasing 1,026 products

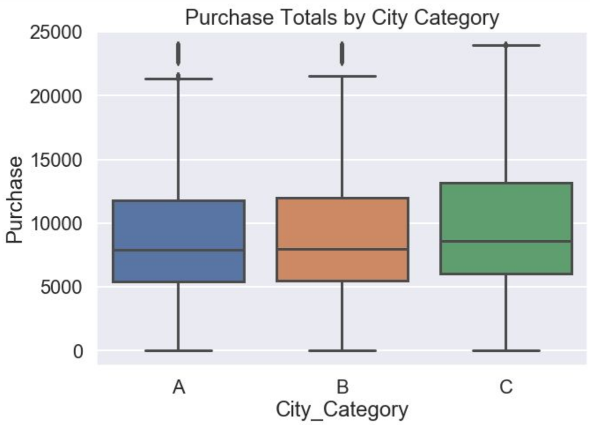


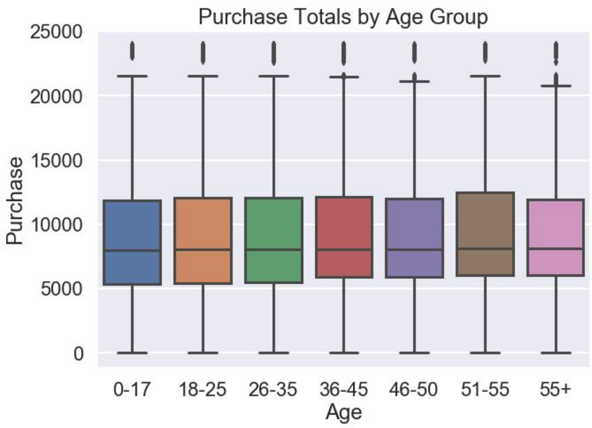
```
count    5891.000000
mean       93.374300
std       107.190049
min         6.000000
25%        26.000000
50%        54.000000
75%       117.000000
max      1026.000000
```

There is not a large difference in average purchase amounts among the customer demographic information.

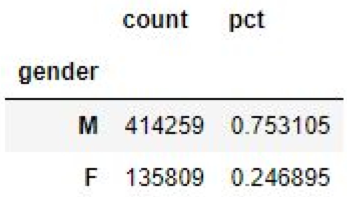






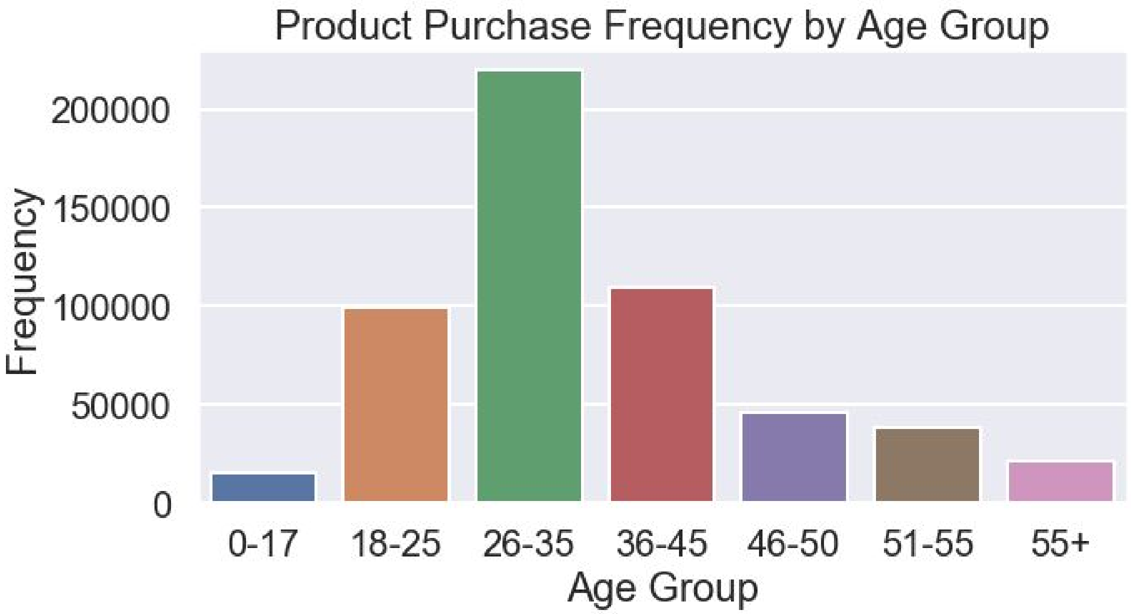


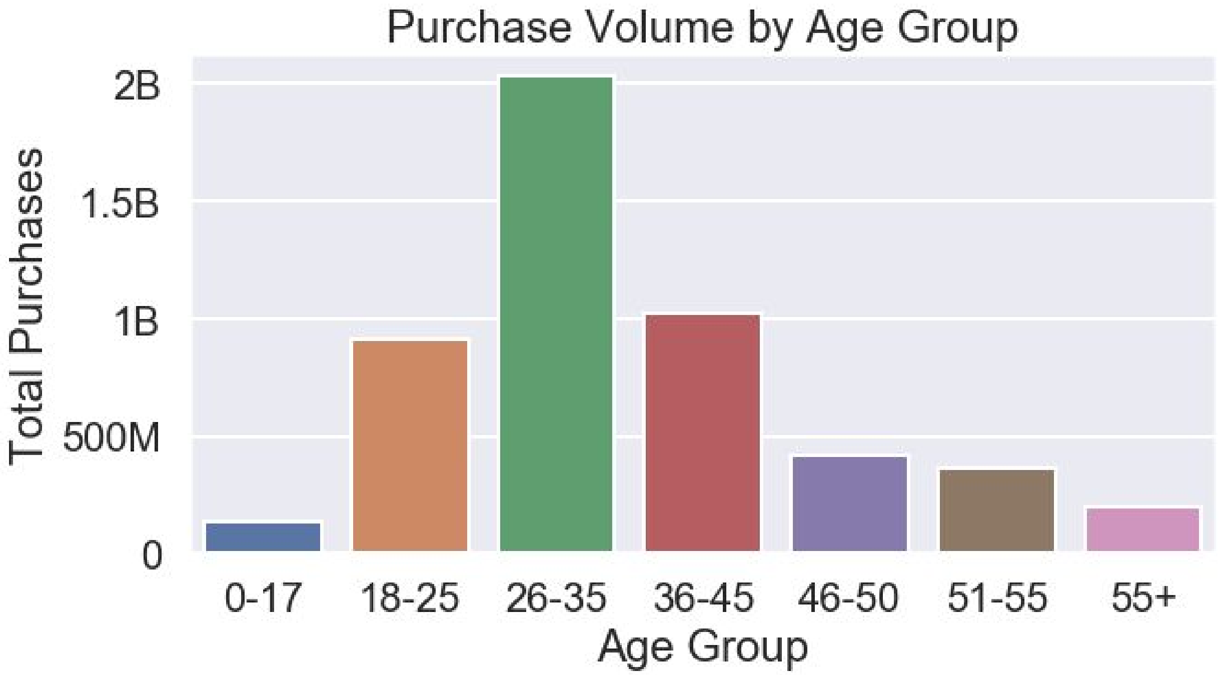
There was a large observance of male customers making 75% of the purchases



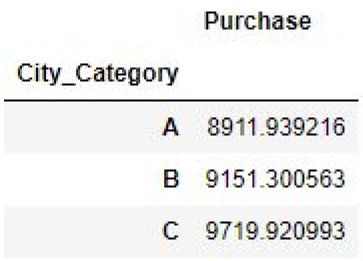
	count	pct
gender		
M	414259	0.753105
F	135809	0.246895

There was a large observance of customers aged 26-35 making 40% of the purchases



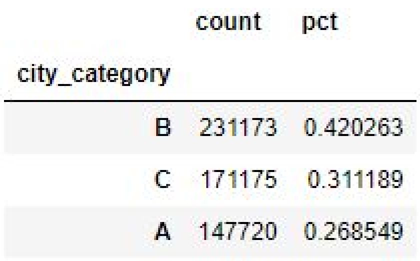


City Category C had the largest spend among the feature



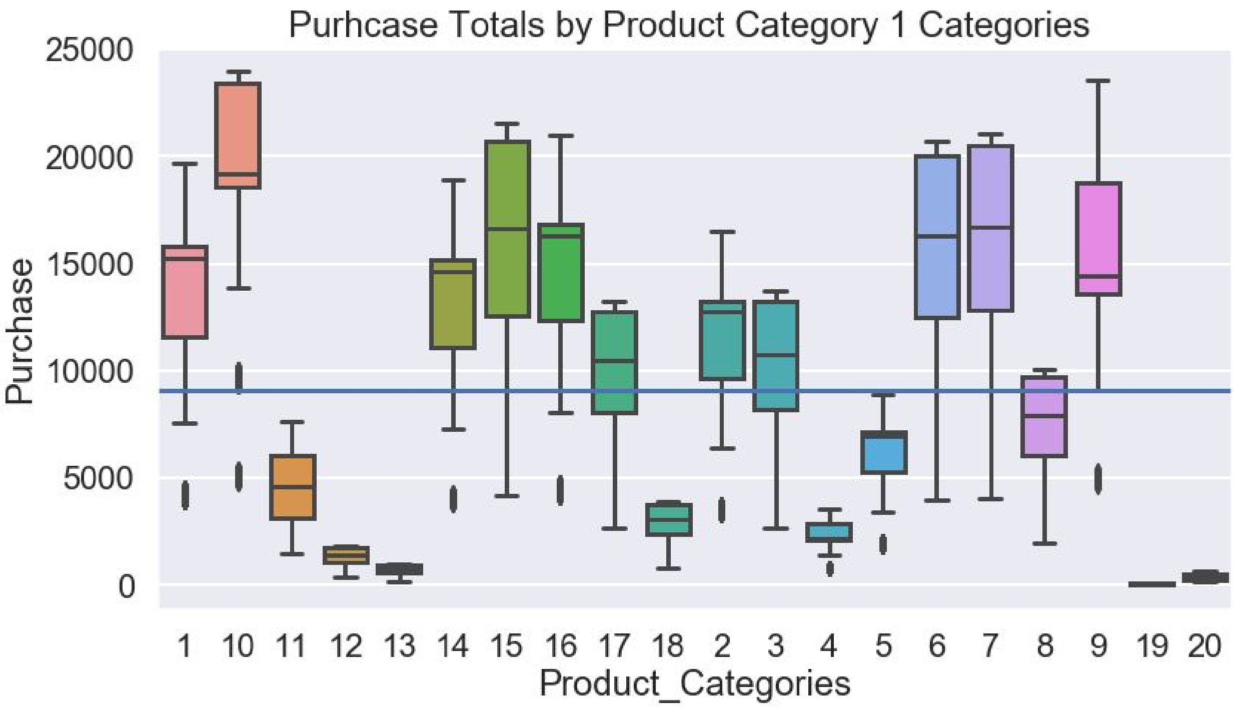
	Purchase
City_Category	
A	8911.939216
B	9151.300563
C	9719.920993

City Category B was the most frequent customer, followed by City_Category C

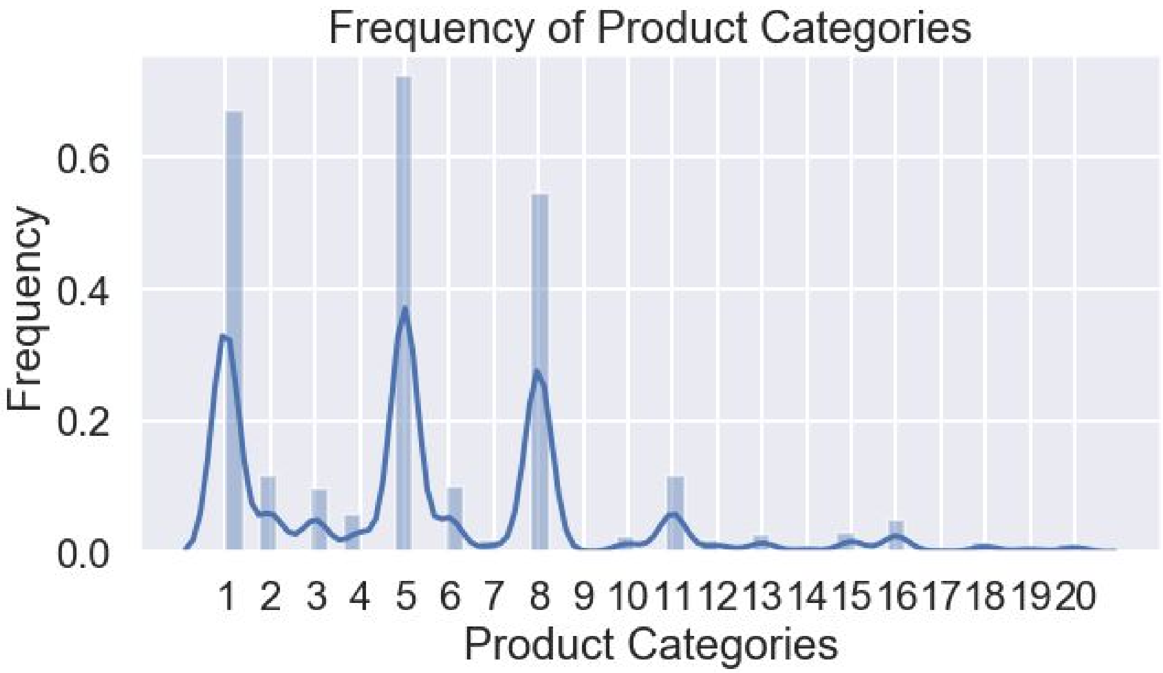


	count	pct
city_category		
B	231173	0.420263
C	171175	0.311189
A	147720	0.268549

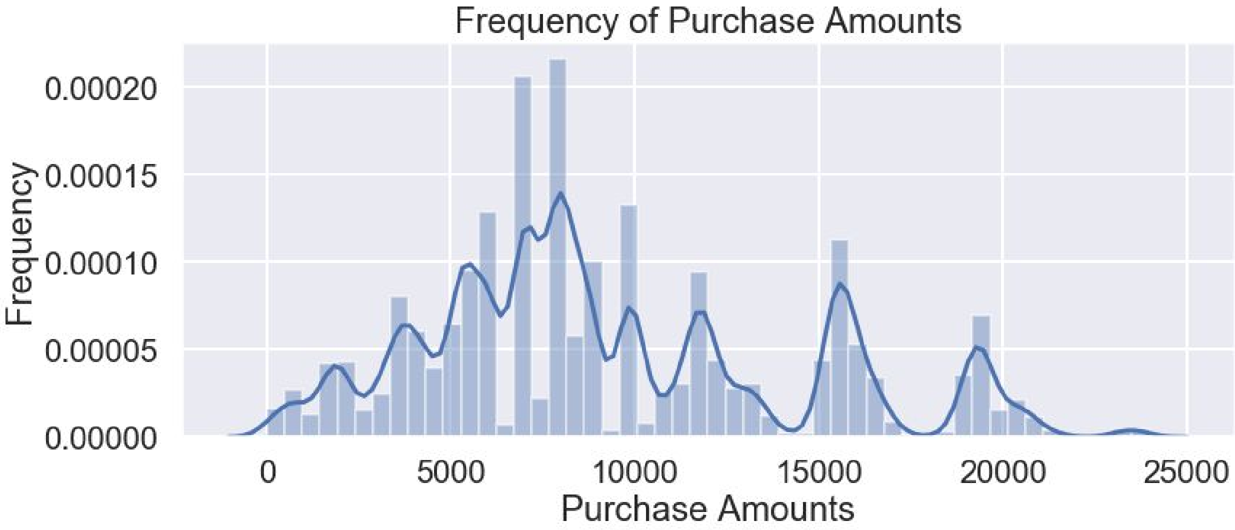
While there was no large variability among customer data when it came to purchase amounts, there was large variability among product categories. Product Categories ranging from $37 to $17,361.



Products 5, 1 and 8 were the dominant product categories among the data. Which is interesting that products 5 and 8 were below the mean of purchases and product 1 was well above. Determining on what these products are, can we shift purchases from 5 or 8 to 1, to drive additional sales?

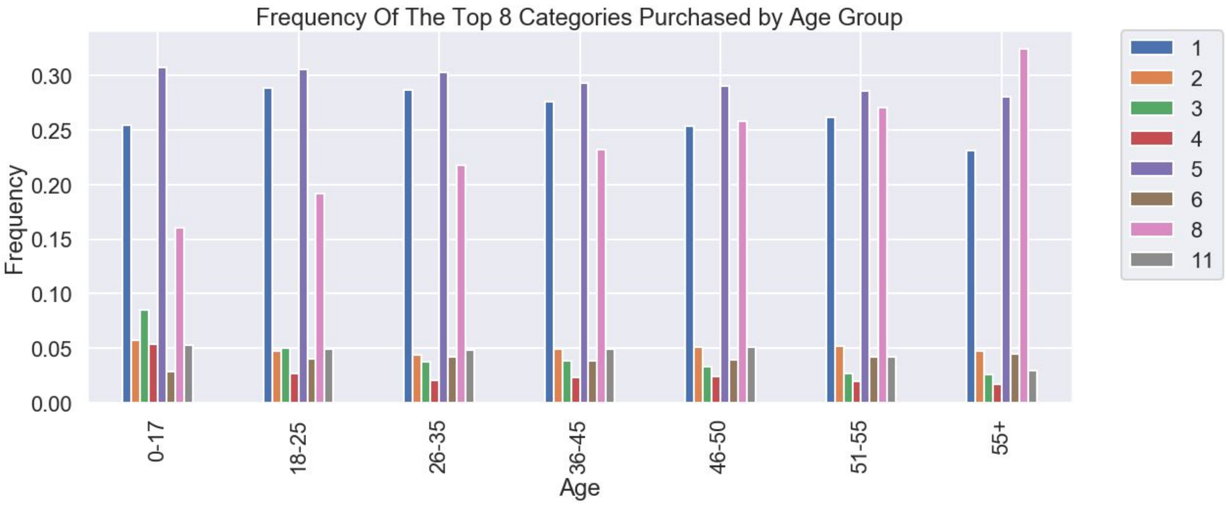


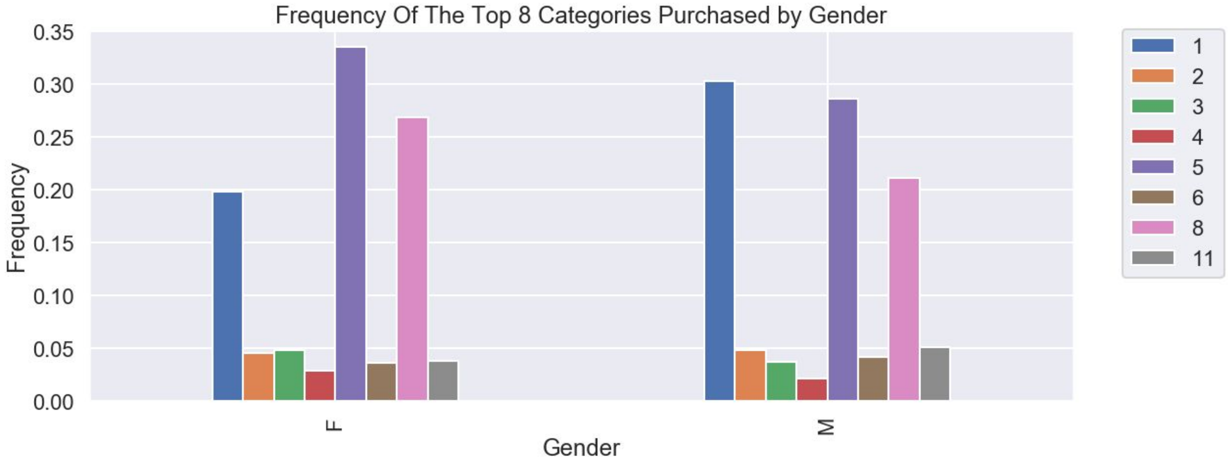
The majority of purchases were less than $15,000 with two peaks at $16,000 and $20,000.

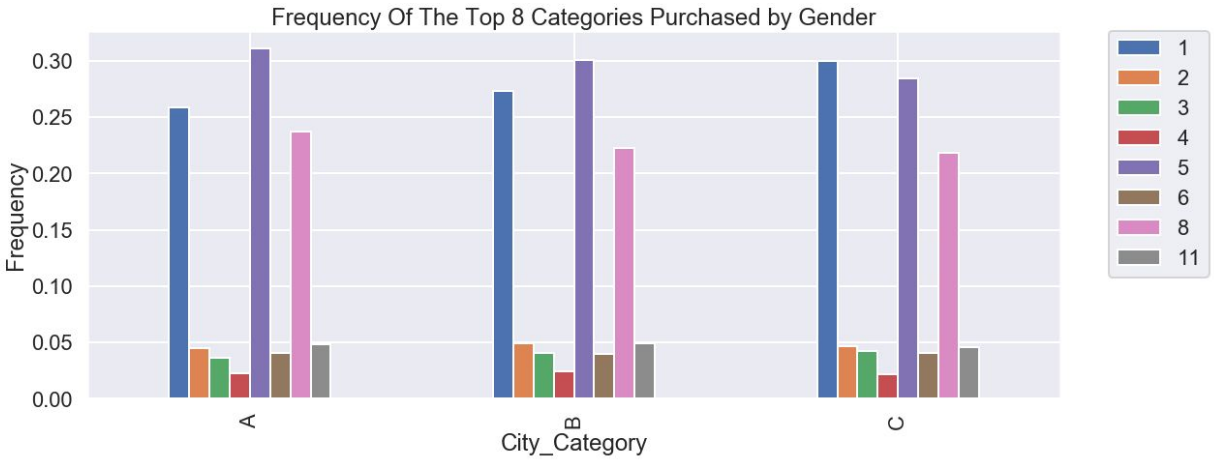


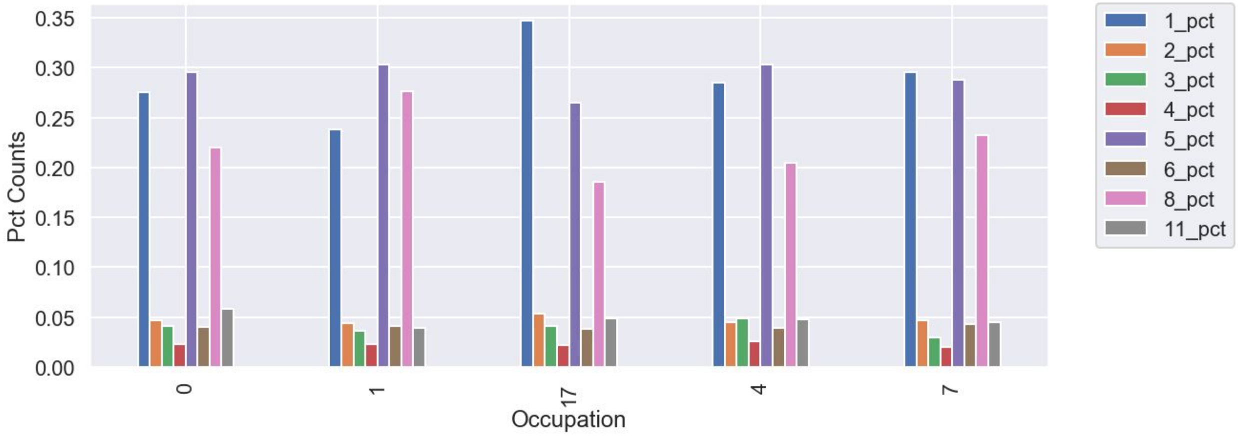
Observing that there is little variability among purchase amounts for particular customer demographic data. We see that there is variability to the proportion of the top volume products by age group. As the age groups increase, Product Category 5 and 1 decrease and Product category 8 increases. We also see that females' top two categories are different than males.

We see similar trends among the City_Category and Occupation features.

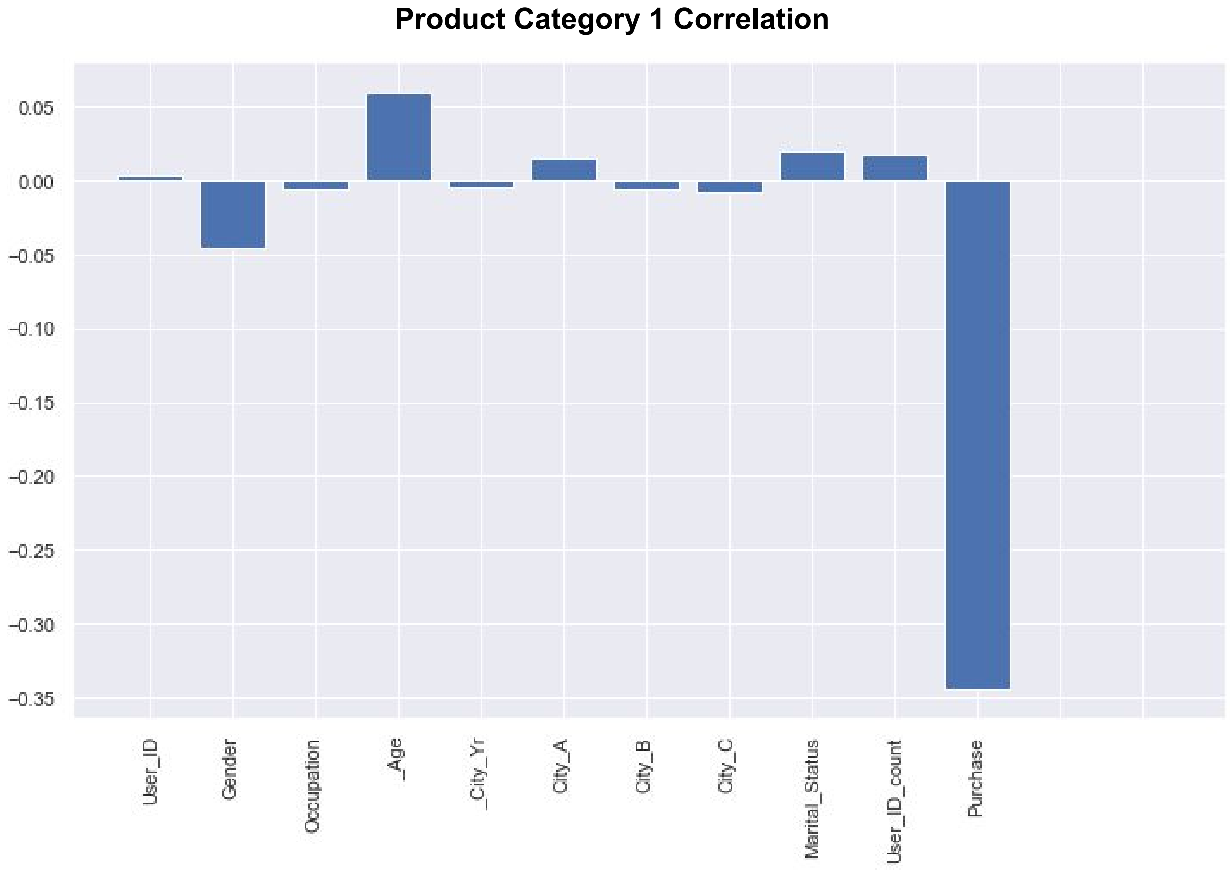








There were no strong correlations between the product category and the other customer demographic features, the features had a pearson correlation of less than .1. This is strange given the charts above. I believe this is due to the number of purchases for each customer. Each observance is one product, and one customer may have multiple purchases of different categories in the data. We saw the average was 93 products, each with possibly different categories.



Product Category 1 Correlation