

**Problem Statement**

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month.

The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offers for customers against different products.

Client

ABC Private Limited

Data Set

The data set can be found at the Analytics Vidhya website. It is the data set for the hackathon site : <https://datahack.analyticsvidhya.com/contest/black-friday/>

**Data**

**Variable** Definition

**User\_ID** User ID

**Product\_ID** Product ID

**Gender** Sex of User

**Age** Age in bins

**Occupation** Occupation (Masked)

**City\_Category** Category of the City (A,B,C)

**Stay\_In\_Current\_City\_Years** Number of years stay in current city

**Marital\_Status** Marital Status

**Product\_Category\_1** Product Category (Masked)

**Product\_Category\_2** Product may belongs to other category also(Masked)

**Product\_Category\_3** Product may belongs to other category also (Masked)

**Purchase** Purchase Amount (Target Variable)

Demographics: The key to this problem

**HYPOTHESIS**

Every year Black Friday offers great discounts to the buyers. This festive bonanza offers a sales peak to the retailers before the festive season peak. They have to stock the goods. This time they can project the amount of sales they could get and the kind of goods they should stock in order to maximize their sales.

The demographics could tell us a lot in the Exploratory Data Analysis

The different age groups, genders, vocations, and stay time in cities can tell us about the buying patterns for these different categories.

We can also see which city has more consumption and why.

The retailers in each city would study the demographic profile of each city. They may start the planning months in advance. In case of consumer durables they might even try to buy the goods in a price trough. This will maximize their per unit cost as they give heavy discounts on the day of black friday.

By building a regressor this problem can be addressed.

These are the correlations we could get.

Data Wrangling

The data comes with two files namely the train.csv and test.csv.

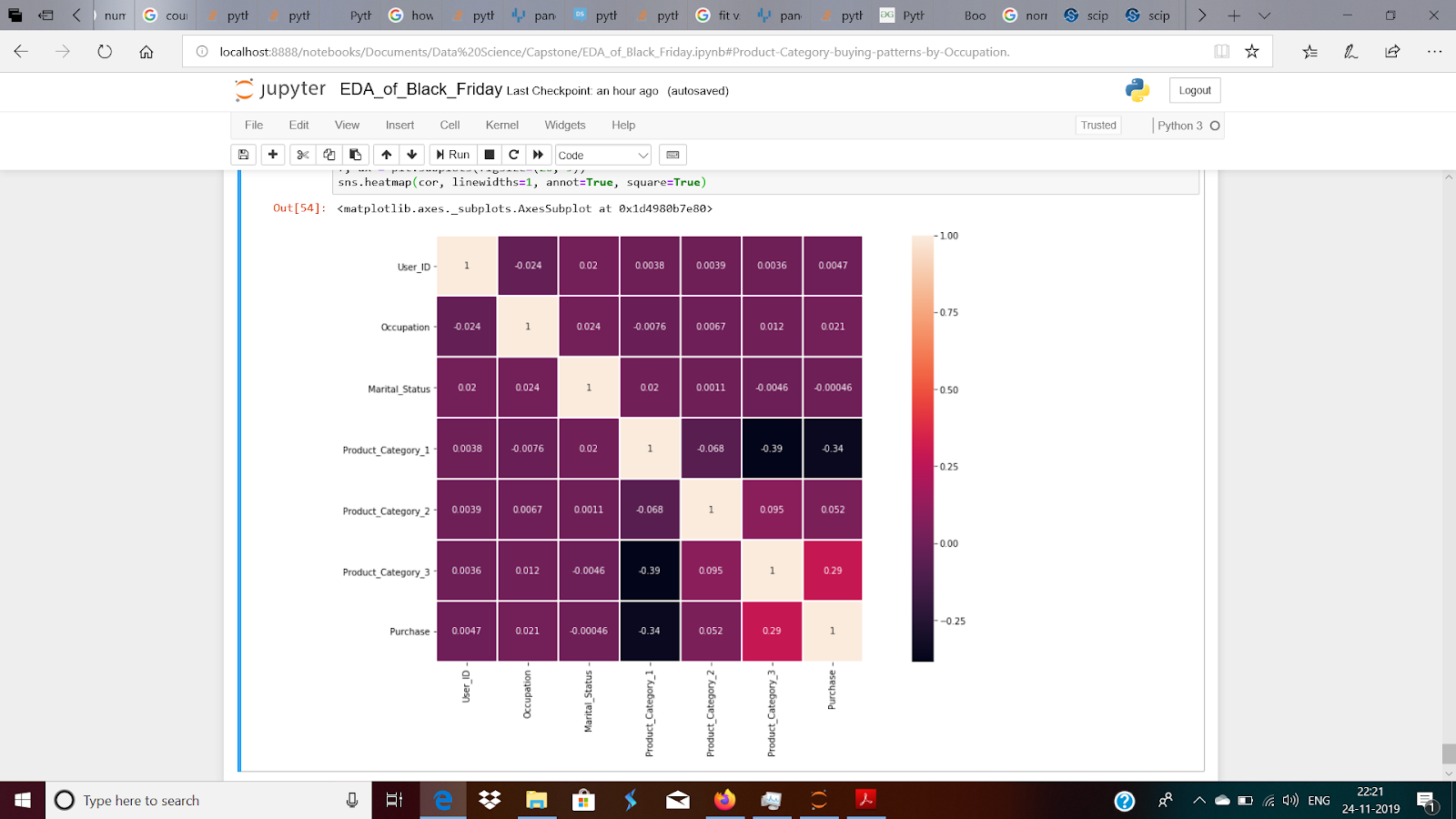
Wrangling is done through the following steps:

* Inspect the columns of both the data sets. The test data has one less column.
* I check the info(), head, and description of the set.
* Check for null values and see if they can be filled.
* The null values were filled by 0s as they mean the product was not purchased.
* There are duplicates in User IDs but all of those observations have different purchases so we cannot delete the duplicates.
* Moreover the same product is sold at a different price in different transactions
* The dataset has 12 columns or features

Exploratory Data Analysis

Exploratory Data Analysis was carried out by constructing graphical representations of the various features.

* Countplots, violinplots were used to see the relationship between different features.
* A correlation heatmap has also been constructed to see the relationship directly in numbers.
* Data types were not needed to be converted to make the graphical representations.
* Some outliers were also observed, but they were not removed because in monetary transactions purchases can be made of any scope on the high level. No negative ‘purchases’ were found.



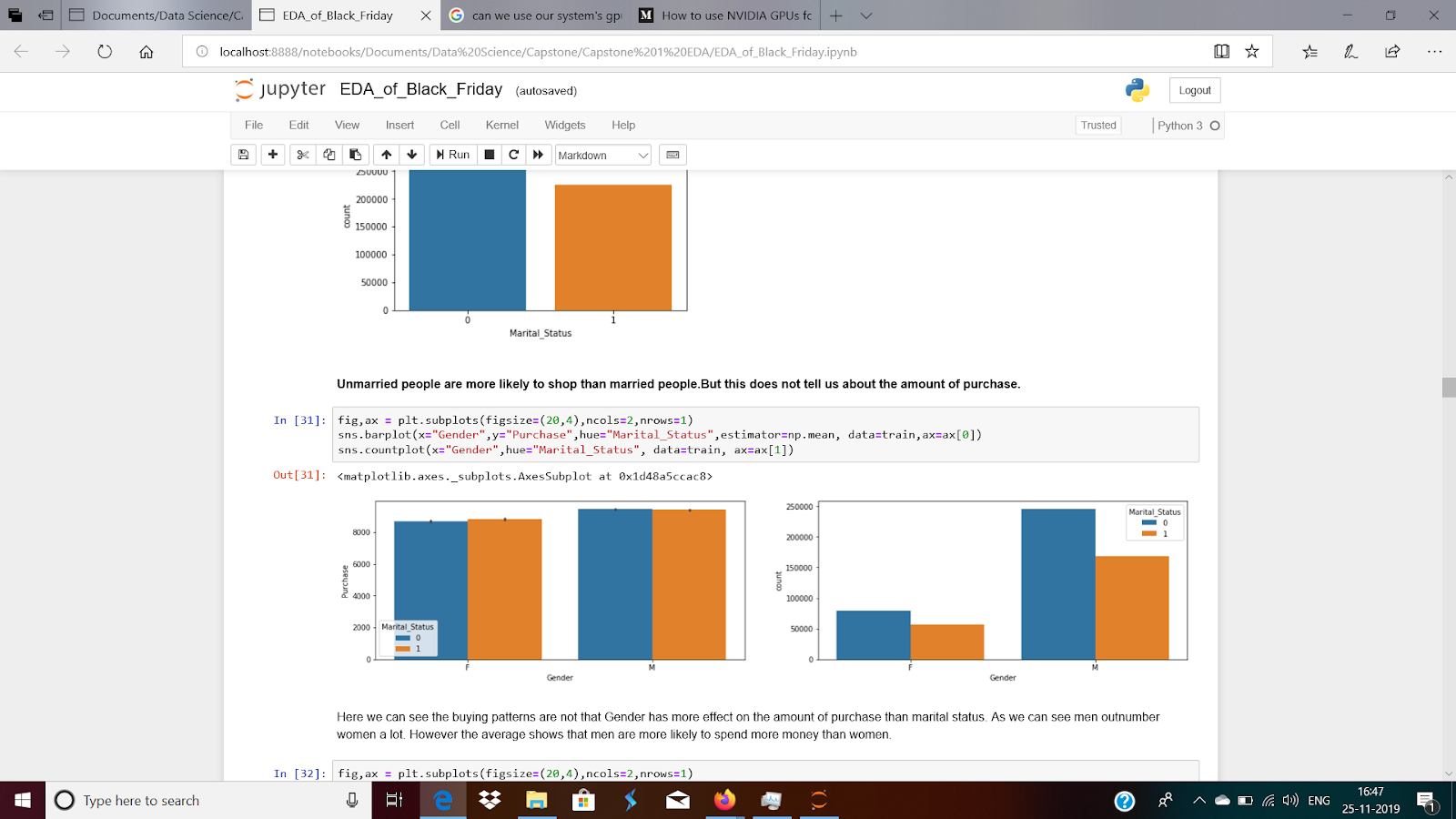
The above correlation heatmap shows that there’s no multicollinearity.

Visual Exploratory Data Analysis

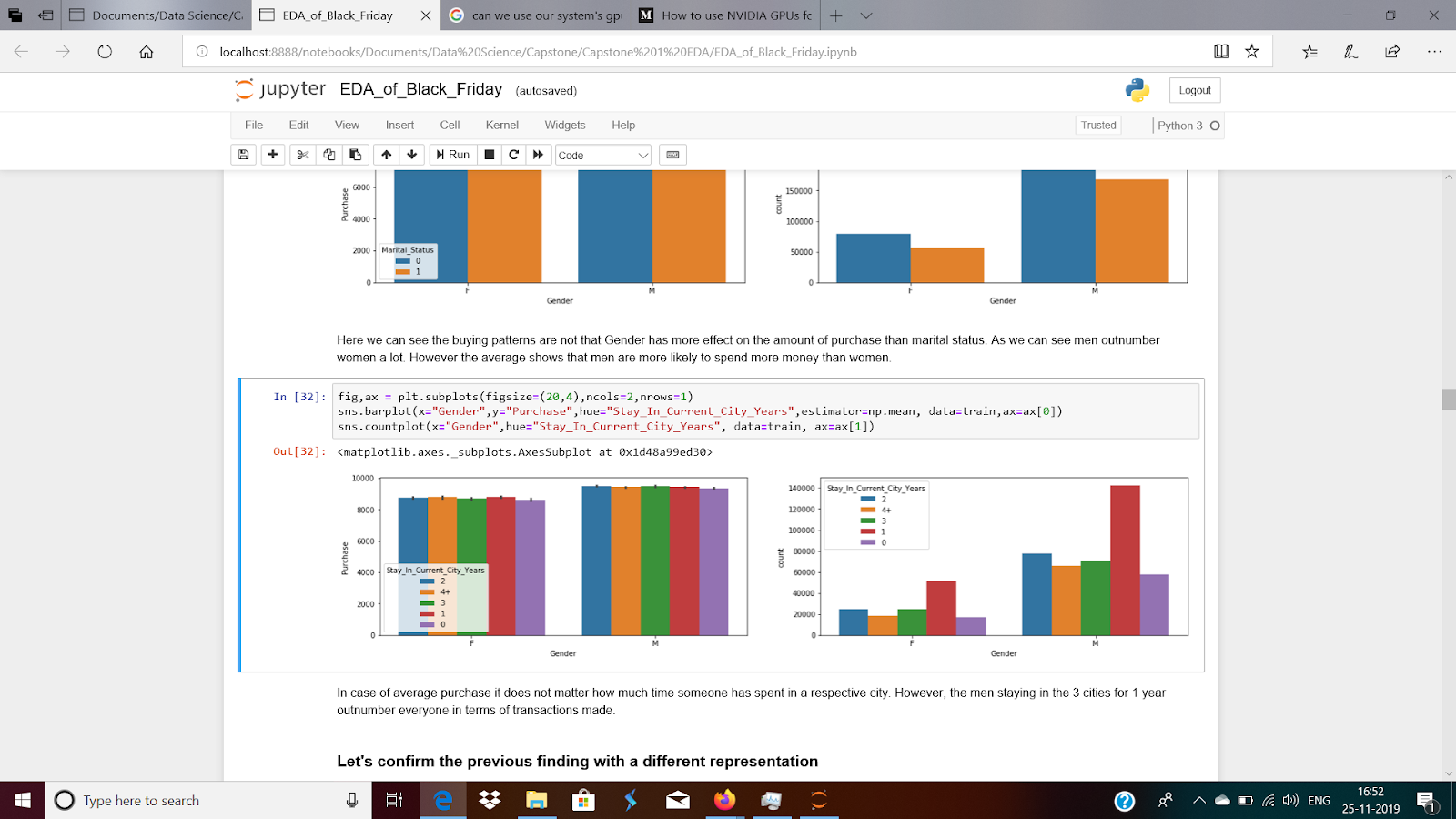
Initial Findings

During exploratory data analysis, we ask the following questions:

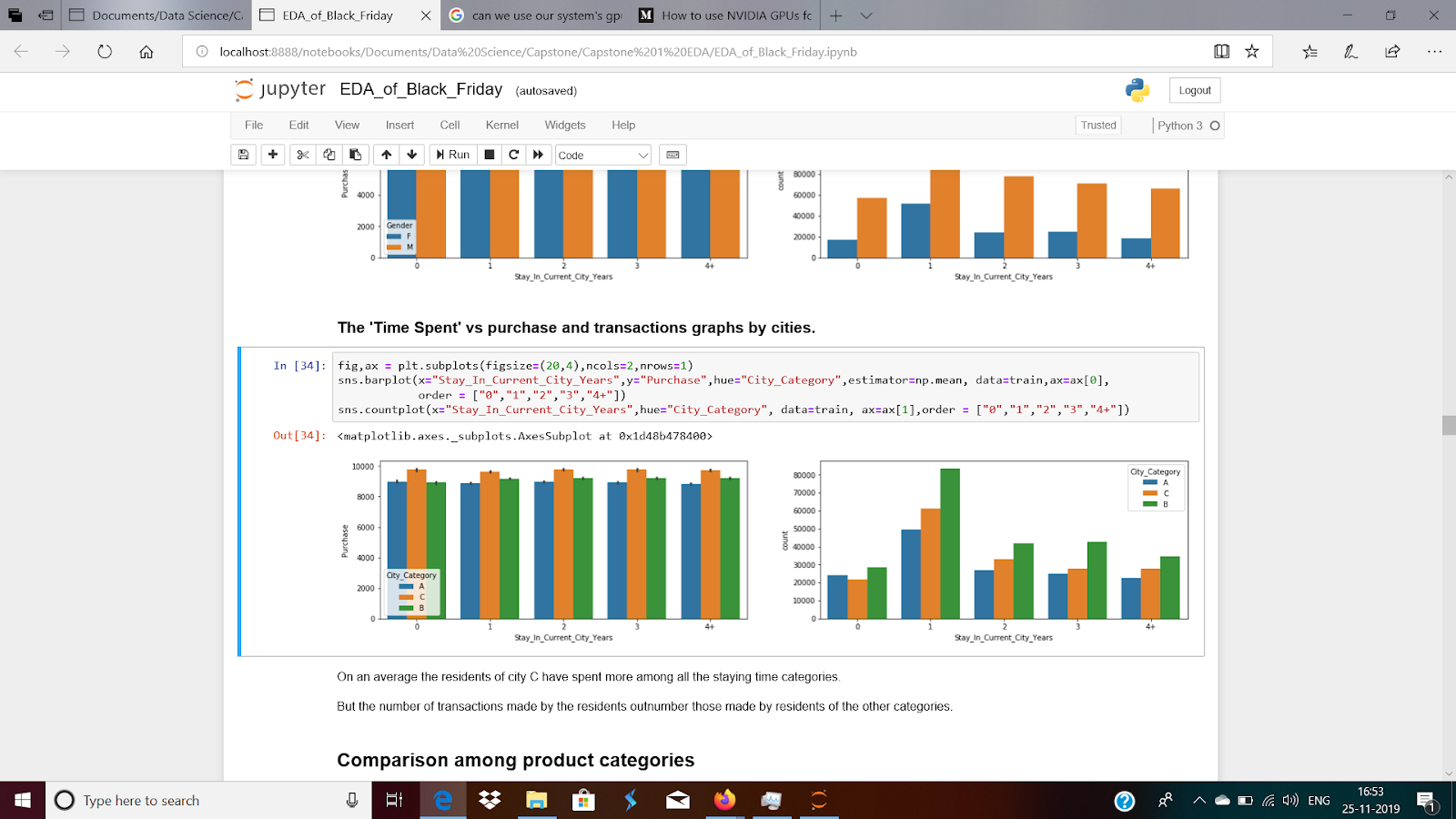
* What might be the most important features that affect listing price?
* Does city affect listing price?
* Does the marital status or



Here we can see the buying patterns are not that Gender has more effect on the amount of purchase than marital status. As we can see men outnumber women a lot. However the average shows that men are more likely to spend more money than women.

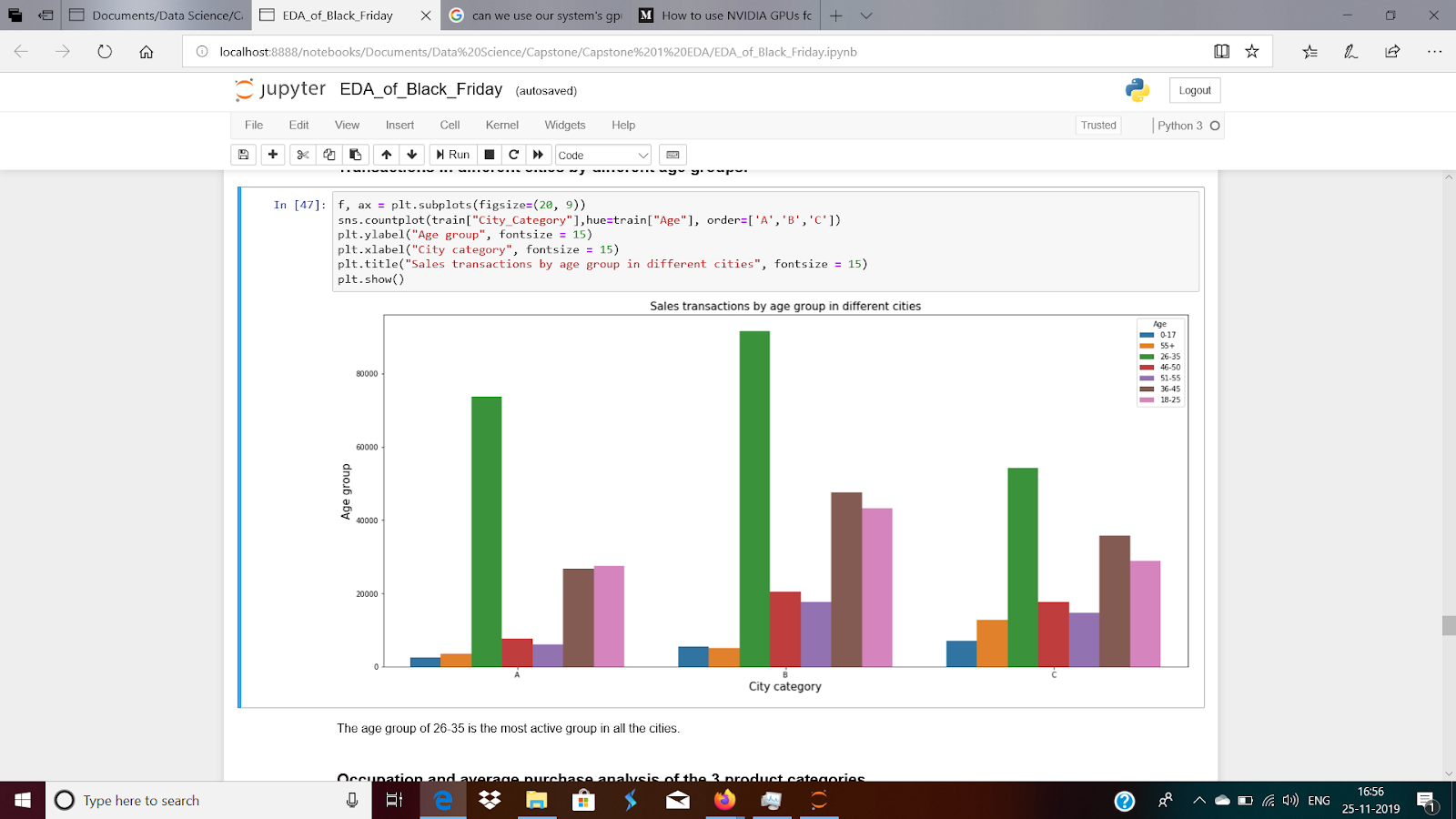


In case of average purchase it does not matter how much time someone has spent in a respective city. However, the men staying in the 3 cities for 1 year outnumber everyone in terms of transactions made.

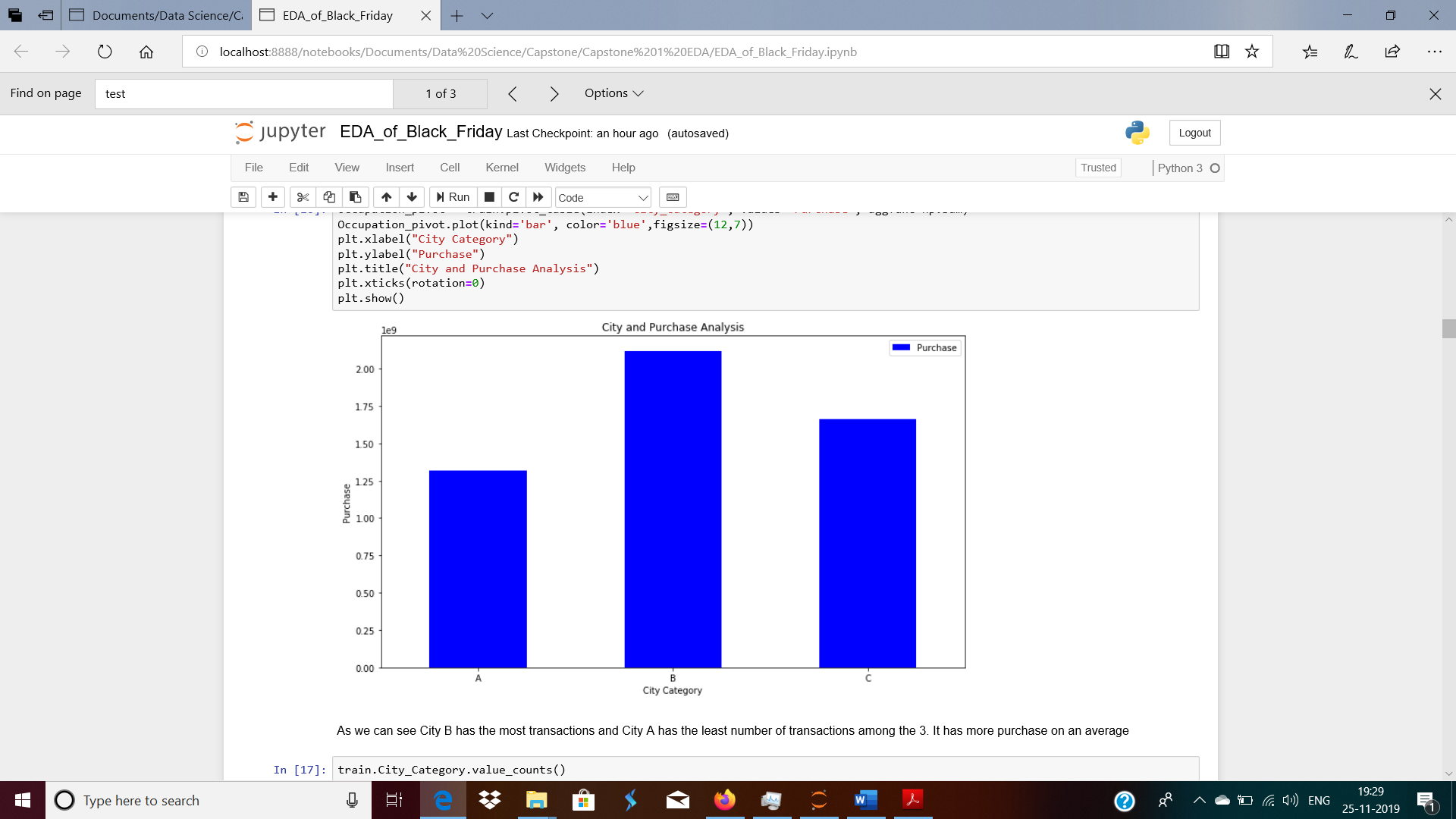


On an average the residents of city C have spent more among all the staying time categories.

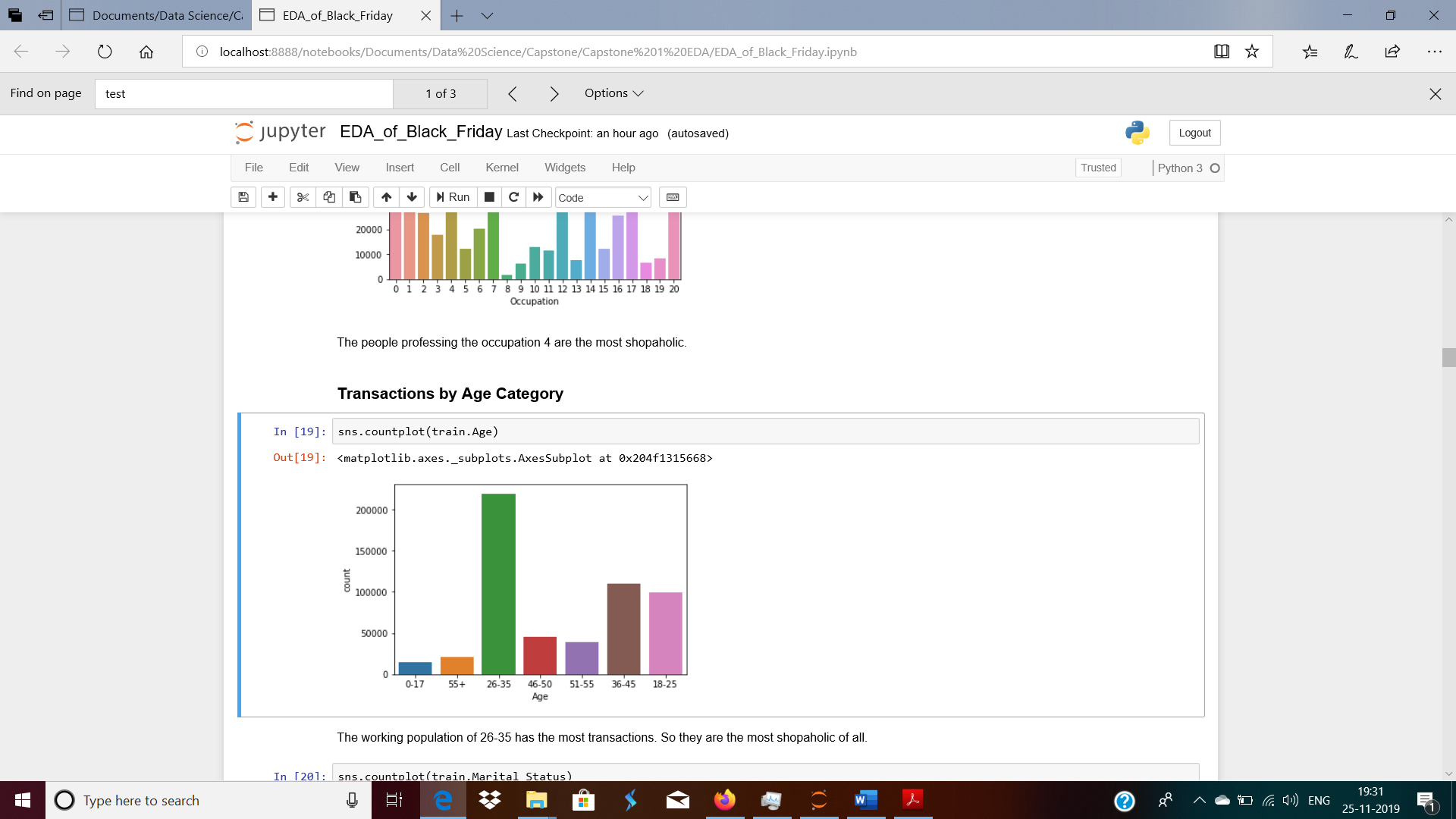
But the number of transactions made by the residents outnumber those made by residents of the other categories.



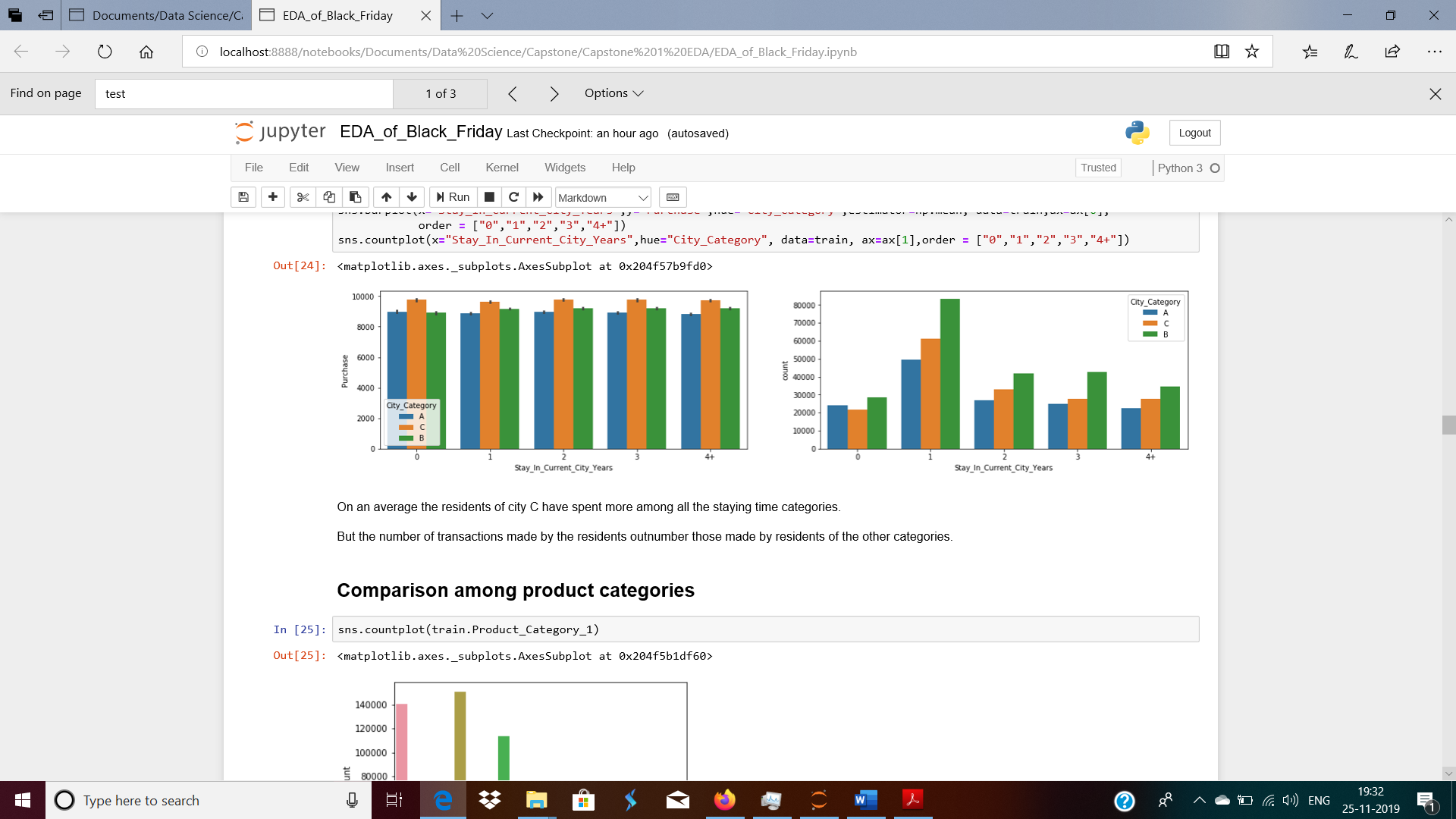
The age group of 26-35 is the most active group in all the cities.



As we can see City B has the most transactions and City A has the least number of transactions among the 3. It has more purchase on an average



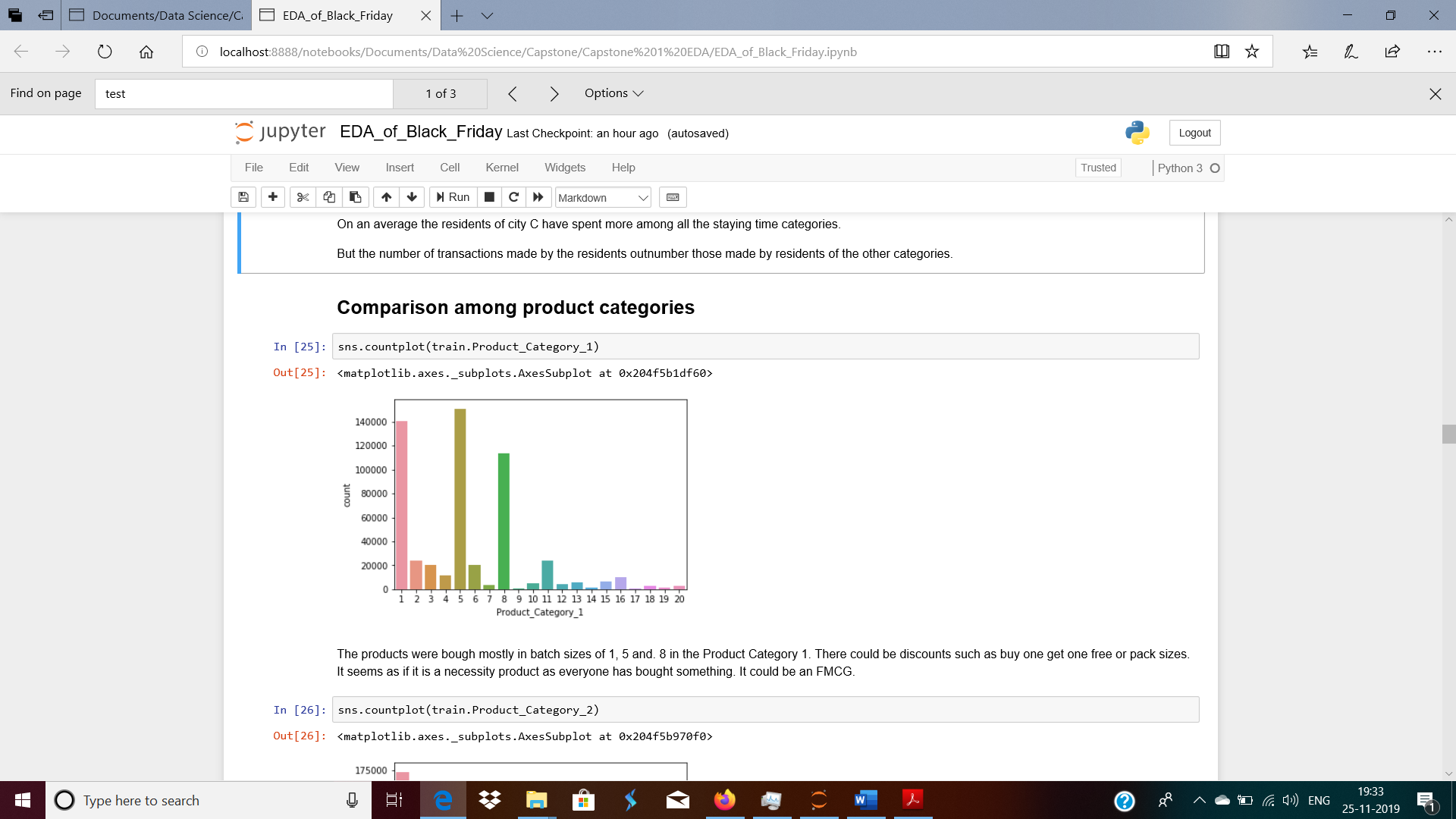
The working population of 26-35 has the most transactions. So they are the most shopaholic of all.



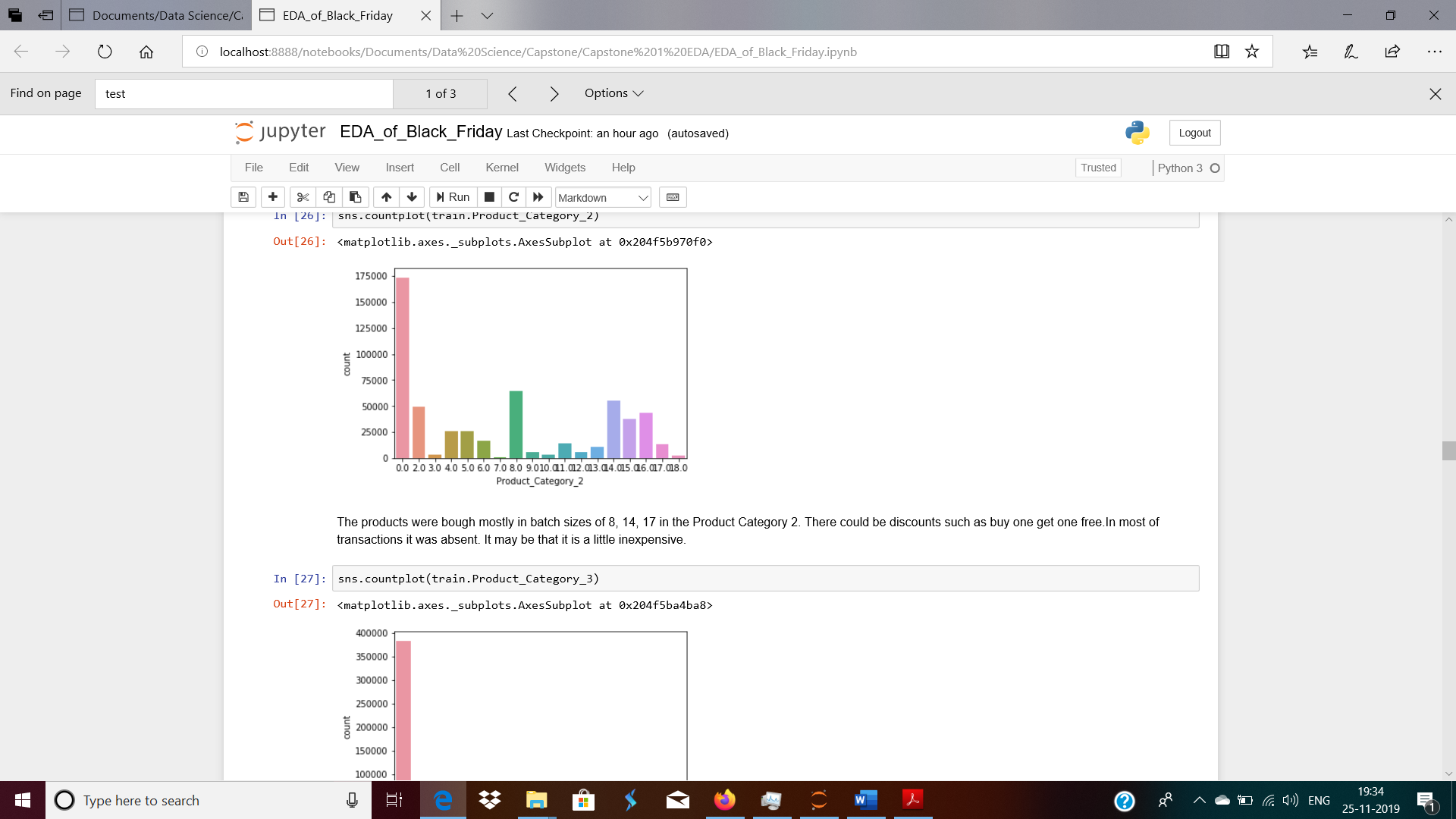
On an average the residents of city C have spent more among all the staying time categories.

But the number of transactions made by the residents outnumber those made by residents of the other categories.

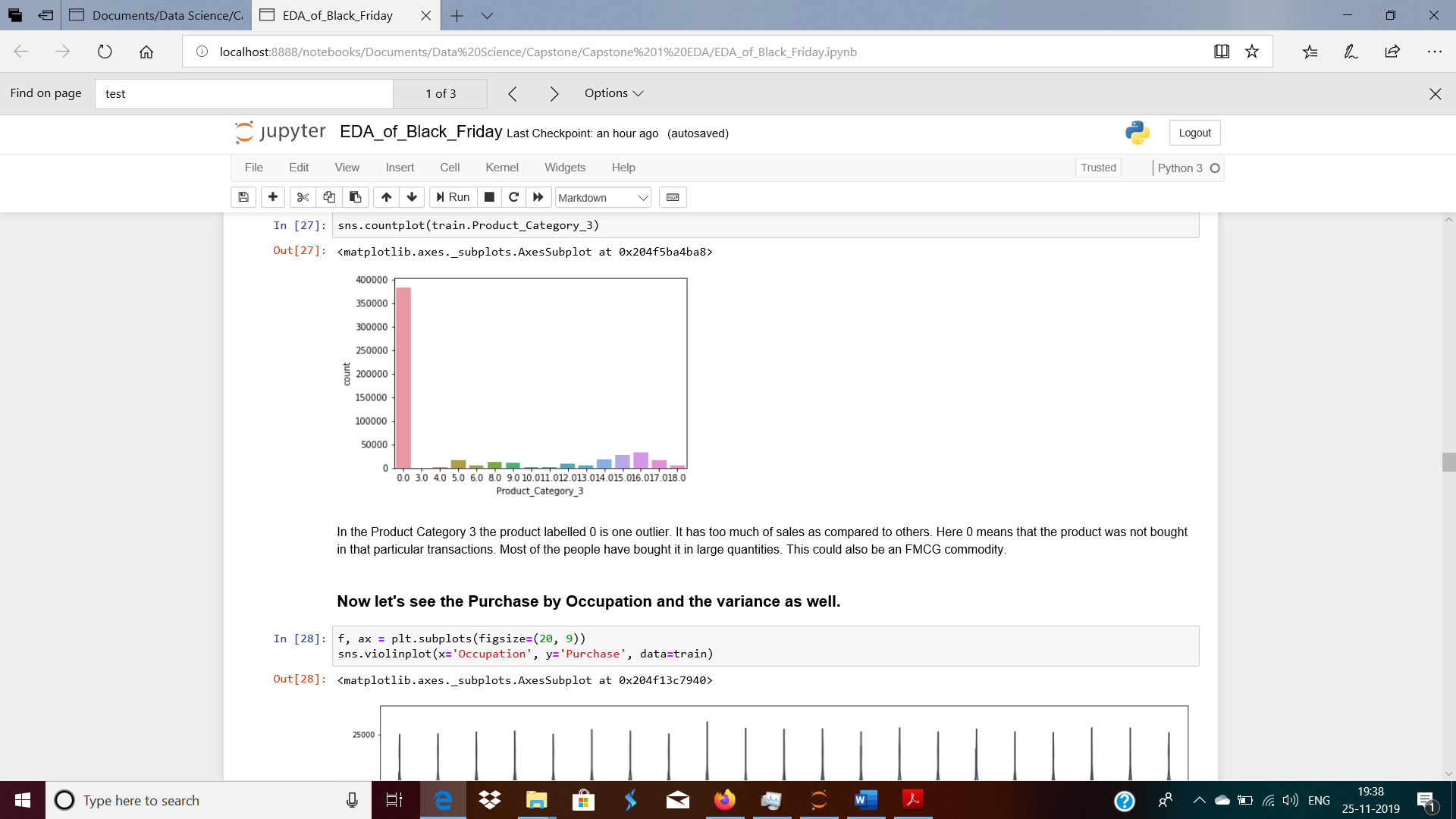
The products were bough mostly in batch sizes of 1, 5 and. 8 in the Product Category 1. There could be discounts such as buy one get one free or pack sizes. It seems as if it is a necessity product as everyone has bought something. It could be an FMCG.



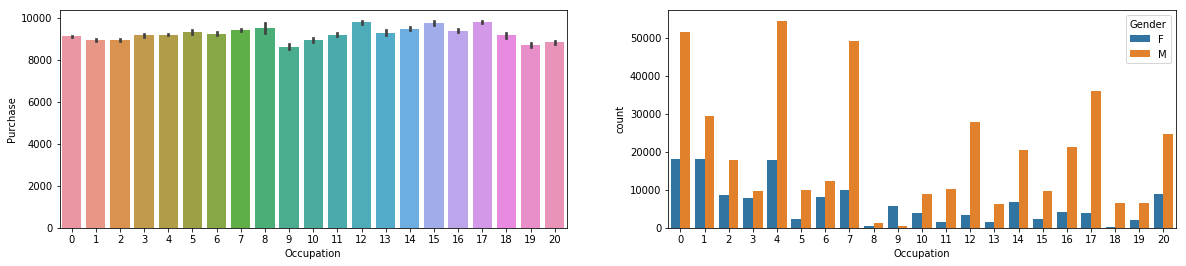
The products were bough mostly in batch sizes of 1, 5 and. 8 in the Product Category 1. There could be discounts such as buy one get one free or pack sizes. It seems as if it is a necessity product as everyone has bought something. It could be an FMCG.



The products were bough mostly in batch sizes of 8, 14, 17 in the Product Category 2. There could be discounts such as buy one get one free.In most of transactions it was absent. It may be that it is a little inexpensive.



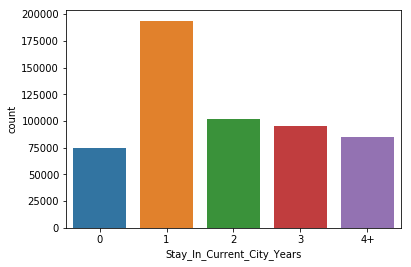
In the Product Category 3 the product labelled 0 is one outlier. It has too much of sales as compared to others. Here 0 means that the product was not bought in that particular transactions. Most of the people have bought it in large quantities. This could also be an FMCG commodity.



We can see on an average the people engaged in the occupation number 9 are the most frugal ones in the lot.

And, the ones in the occupation number 12 are the most spendthrift of the group.

Also in the graph on the right it clearly shows that all of the jobs are dominated by males.



Most of the people in the cities have spent only a single year.

Feature Engineering

For modelling we cannot straightaway go and fit the models. Many of the categorical variables are of object data type that’s why they need to be **encoded first.**

Now the issue was which kind of encoding to go for. I chose **Label Encoding** as most of the features can be ranked ordinally.

## Modelling

With insights based on exploratory data analysis (EDA), we start to train predictive models.

We identified three models to try first:

1. Linear Regression

* Predict the label of a data point by a linear function
* Loss function = Ordinary least squares (OLS), which is sum of squares of residuals

RMSE: 3775.7493615969393

1. Decision Tree Regressor

* **Decision Tree** is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility.

RMSE: 3099.2566225019336

1. Random Forest Regressor

* A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.

RMSE: 2331.24099030193

So ultimately, we chose the Random Forest Regressor.

## Cross Validation

The results are : 0.66920625 0.66959225 0.67003054 0.66702346 0.69556056

The average is : 0.674282612534217

Summary

* Data Wrangling: We checked for all data types and also filled all the null values as 0.
* Exploratory Data Analysis showed us different kinds of relationships among different features.
* We tried 3 different models for sales prediction.
  + - 1. Linear Regression
      2. Decision Tree Regressor
      3. Random Forest Regressor

Random Forest Regressor is the best predictor.