

**Institute of Technology of Cambodia**

**Department of Information and Communication Engineering**

**Subject: Research Methodology**

**Report Assignment-3**

**Professor: SOK Kimheng**

**Student:** YORN Vanda

**ID**: e20181287

**Group**: I5-C

2022-2023

**Report**

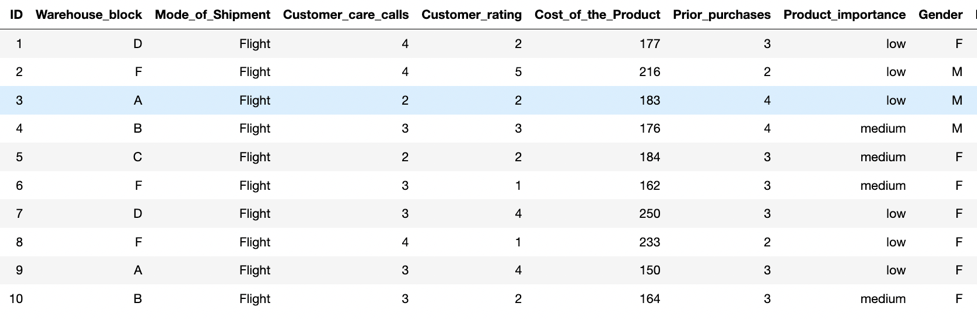
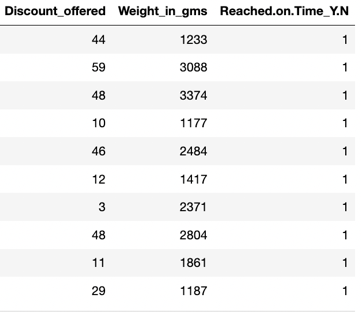
* **About Dataset**

Our dataset describes an international e-commerce company based wants to discover key insights from their customer database. The owner of this dataset wants to use some of the most advanced machine-learning techniques to study their customers. The company sells electronic products.

The data points in this dataset contained 10999 observations of 12 variables. The data contains the following information:

* **ID:** ID Number of Customers.
* **Warehouse block:** The Company have a big Warehouse which is divided into block such as A, B, C, D, E.
* **Mode of shipment:** The Company Ships the products in a multiple way such as Ship, Flight and Road.
* **Customer care calls:** The number of calls made from the enquiry for enquiry of the shipment.
* **Customer rating:** The company has rated from every customer. 1 is the lowest (Worst), and 5 is the highest (Best).
* **Cost of the product:** Cost of the Product in US Dollars.
* **Prior purchases:** The Number of Prior Purchases.
* **Product importance:** The company has categorized the product in the various parameter such as low, medium, high.
* **Gender:** Male and Female.
* **Discount offered:** Discount offered on that specific product.
* **Weight in gms:** It is the weight in grams.
* **Reached on time:** It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

The below figure is showing about the code and libraries that import and show the first 10 data points which are contained in the dataset

****

****

* **Implementation on Statistic Terminology**

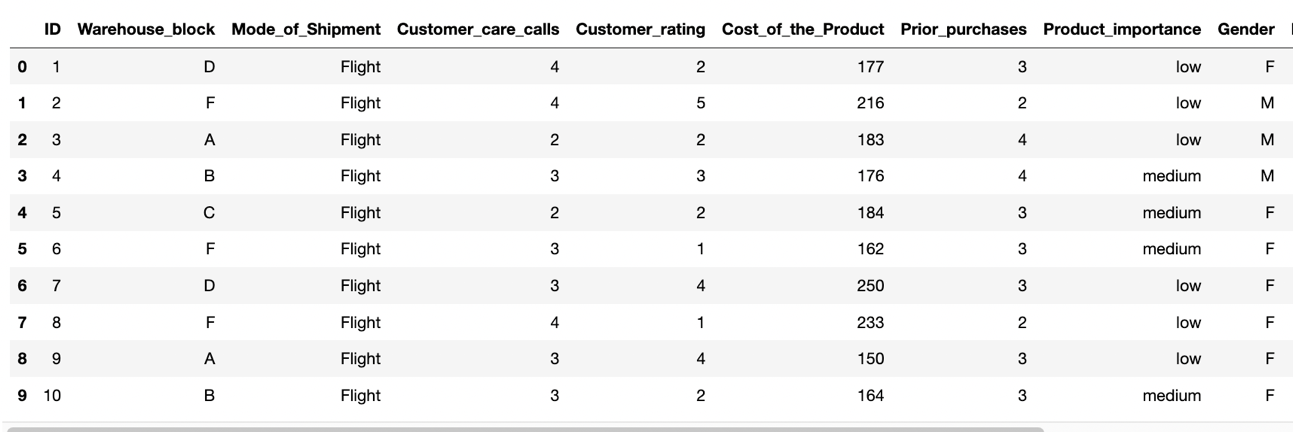
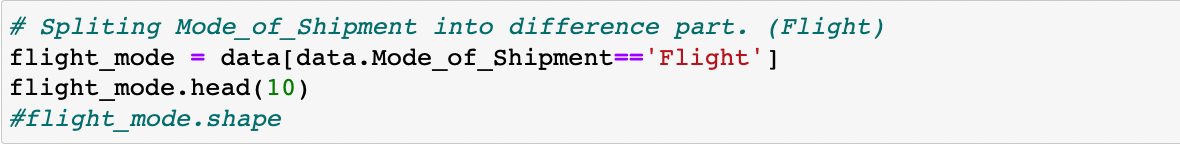
Before go deeper into the dataset, we should know the purpose of the implementation of the dataset.

In the dataset implementation we are going to work on the shipment type, warehouses, and cost of product of the E-Commerce’s shipment.

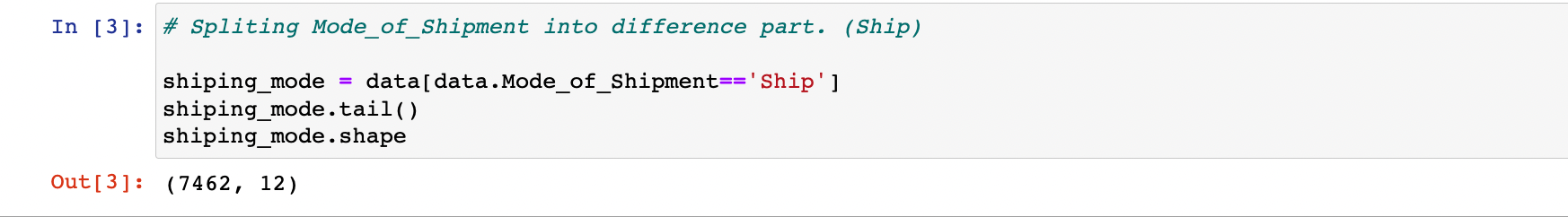
The below descriptive will explain more detail about the implementation:

* **First, I split shipment type into 3 differences**:

~ For Shipment type = Ship

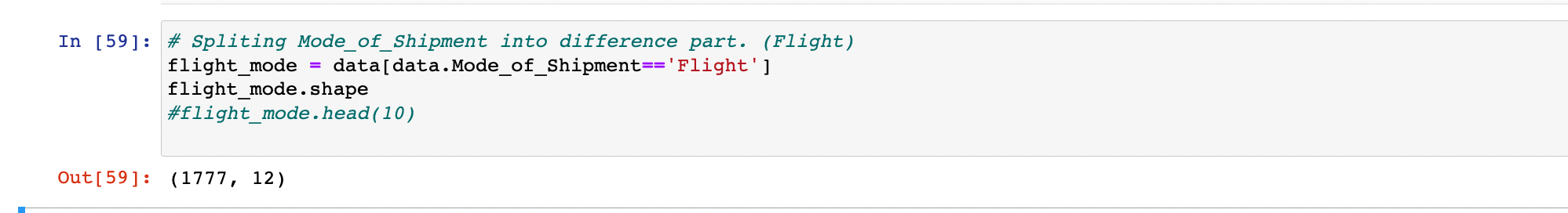


After splitting, we initialized into a new dataset which contained 7462 observations of 12 variables:

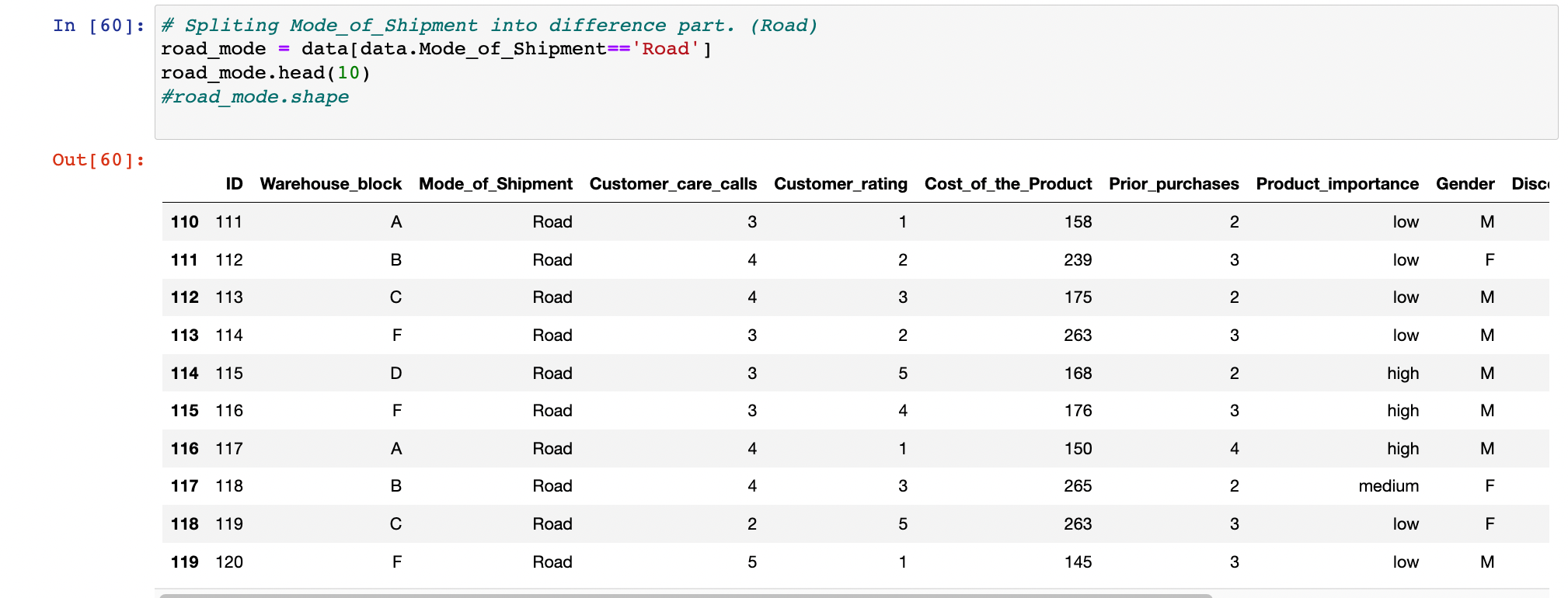


~ For Shipment type = Flight

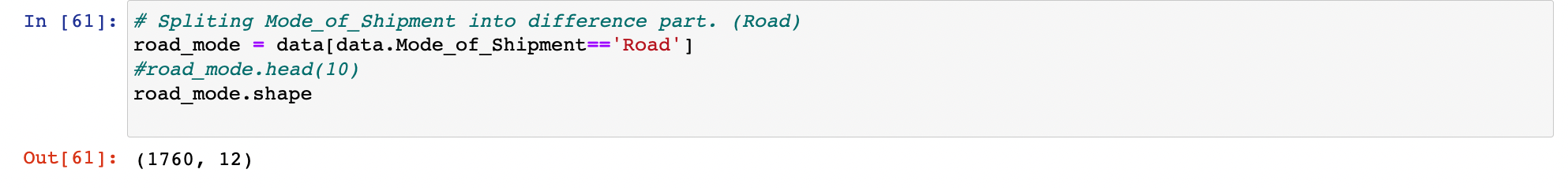
After splitting, we initialized into a new dataset which contained 1777 observations of 12 variables:



~ For Shipment type = Road



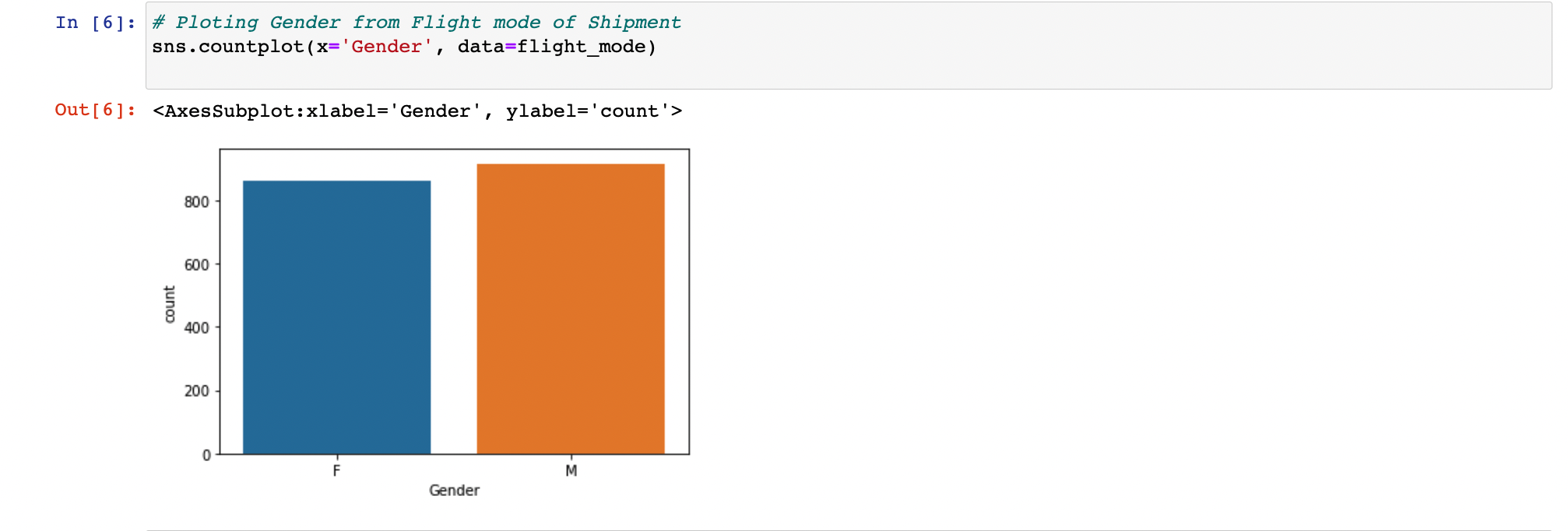
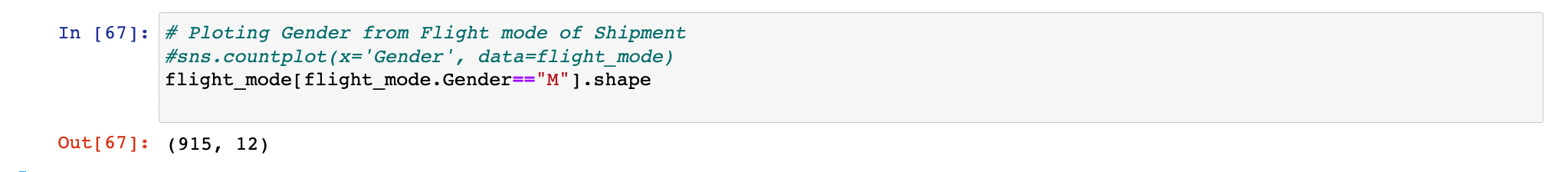
After splitting, we initialized into a new dataset which contained 1760 observations of 12 variables:



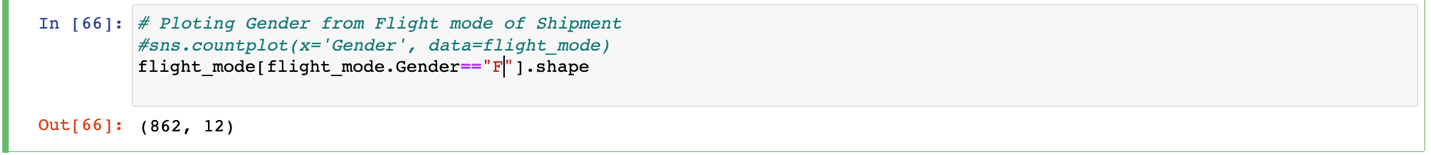
After finished splitting and initialized it into a new dataset, we will use it to do the operation of statistic, it would be easier to access and manage the data than using the whole dataset to do the operation, it could be a bit complicated.

* **Count the number of male and female customers based on Variable “Gender” in each dataset and plot in into bar chart.**

~ For Shipment type = Flight

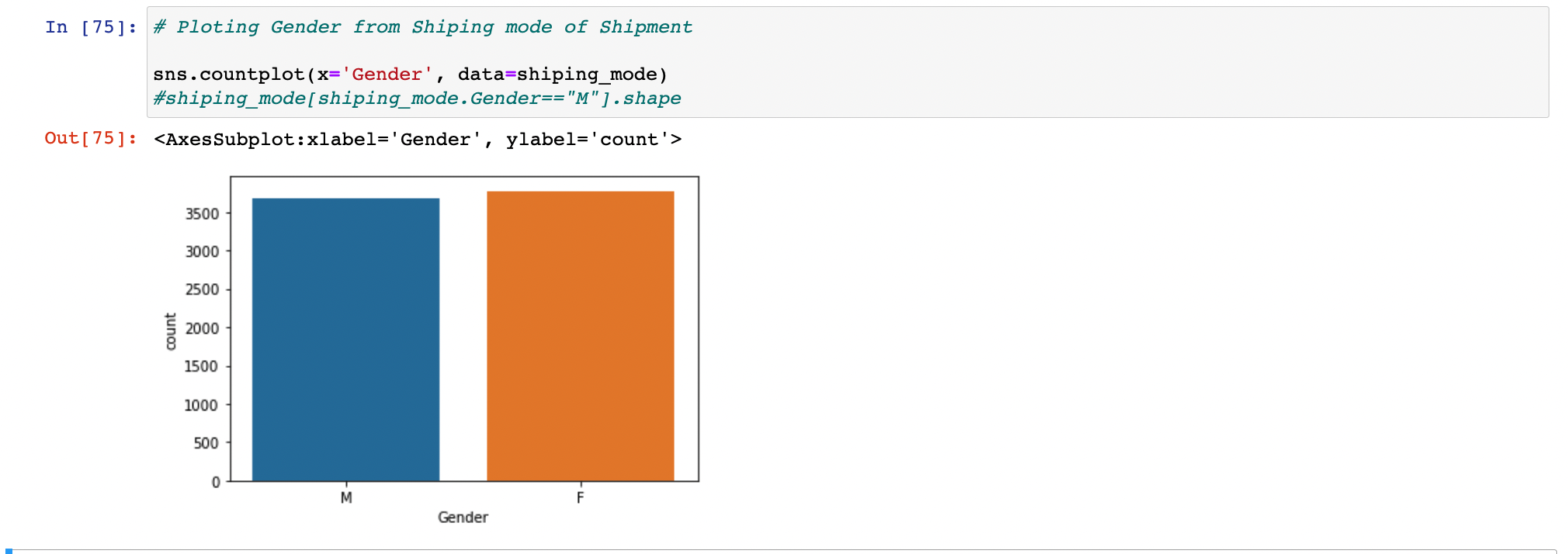
In the new dataset “Flight\_Mode” number of both male and female are1777. In this number we see that number of male customers are more than female 915 and 862, respectively

Male Customers:

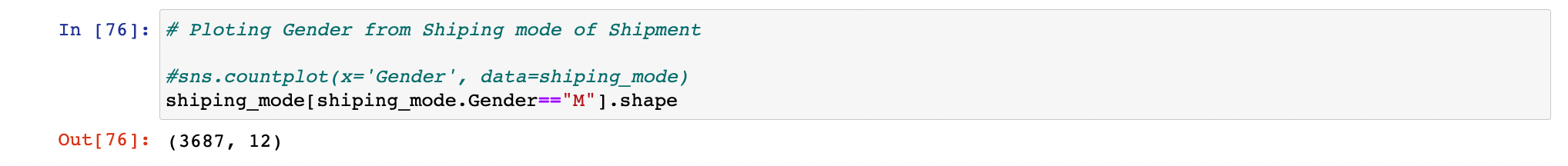
Female Customers:

~ For Shipment type = Ship

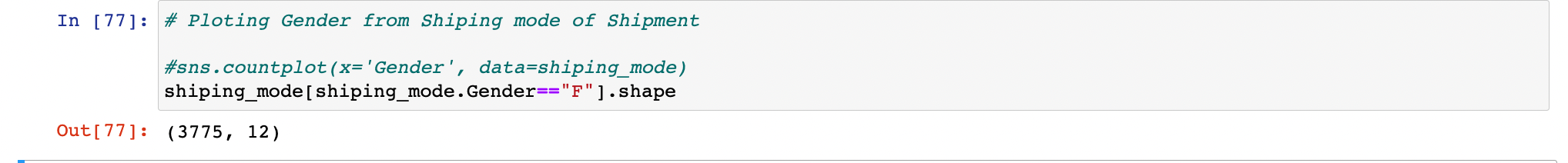
In the new dataset “Shipping\_Mode” number of both male and female are7462. In this number we see that number of female customers are more than male 3775 and 3687, respectively



Male Customers:



Female Customers:

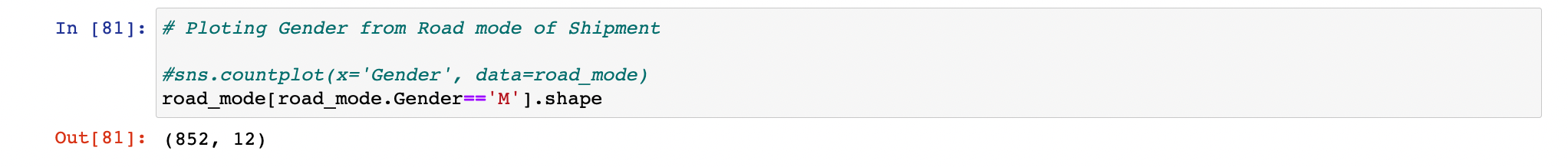


~ For Shipment type = Road

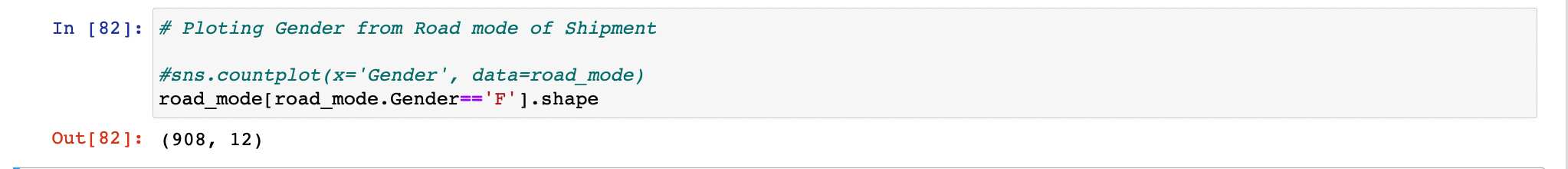
In the new dataset “Road\_Mode” number of both male and female are1760. In this number we see that number of male customers are less than female 852 and 902, respectively



Male Customers:



Female Customers:



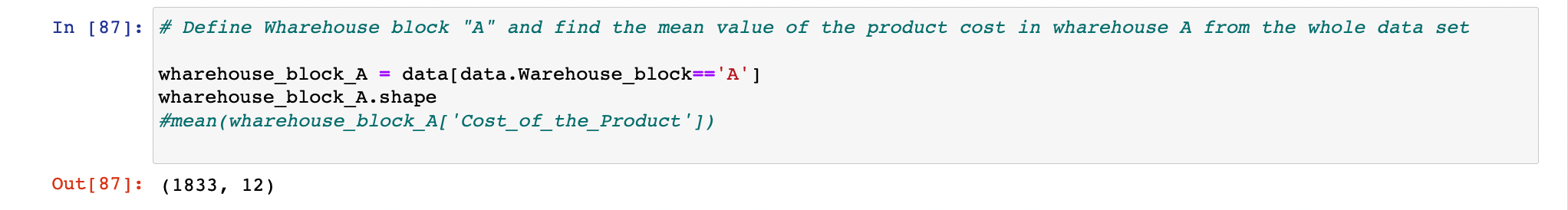
* **Measure of Central tendency (Mean, Median, Mode) of Variable “Cost\_of\_the\_Product”**

Due to the E-Commerce busines concept, the Company have big Warehouse which is divided in to block such as A, B, C, D, E.

Since the main purpose of this topic is only to study about the Statistic Terminologies. Since then we choose only one warehouse to do the operation which contained 1833 data points.

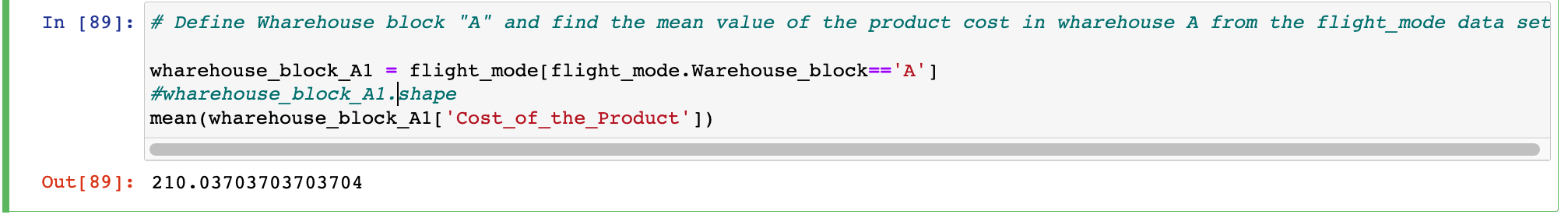
\*\*\* **NOTE:** Only warehouse A which is studied.

~ For all Shipment type of Warehouse A



~ For Shipment type = Flight

* Mean Value = 210.037



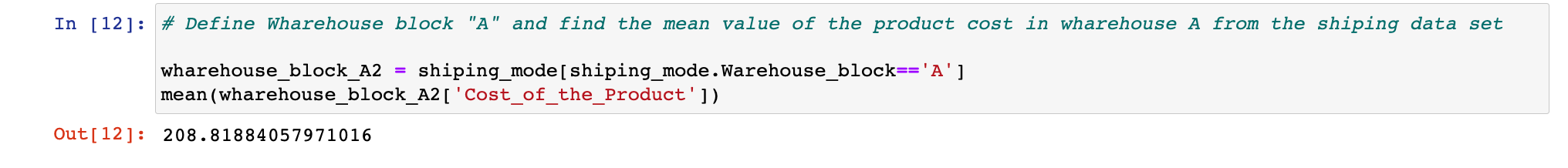
* Median Value = 216

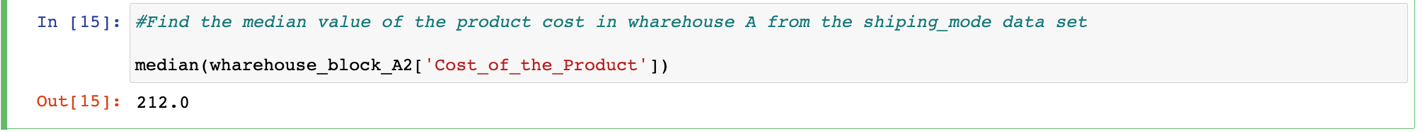


* Mode Value = 255

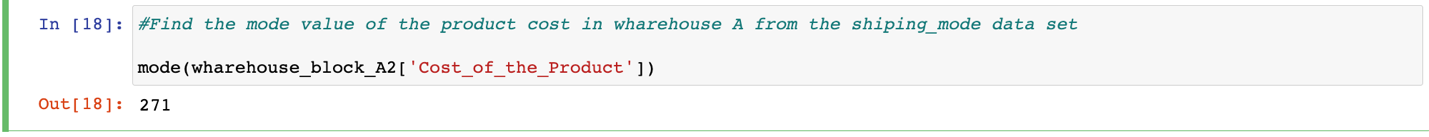


~ For Shipment type = Ship

* Mean Value = 208.818
* Median Value = 212.0

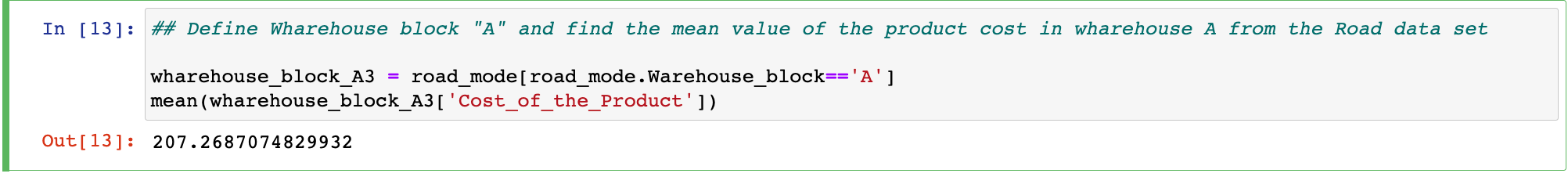


* Mode Value = 271

****

~ For Shipment type = Ship

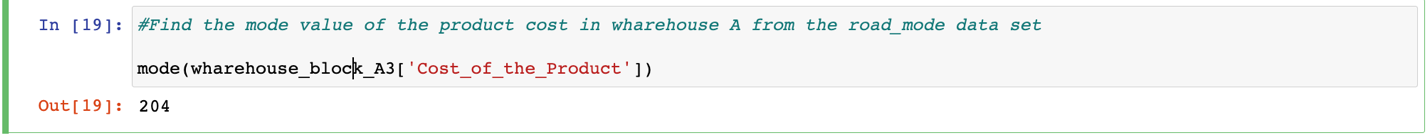
* Mean Value = 207.268



* Median Value = 207.5



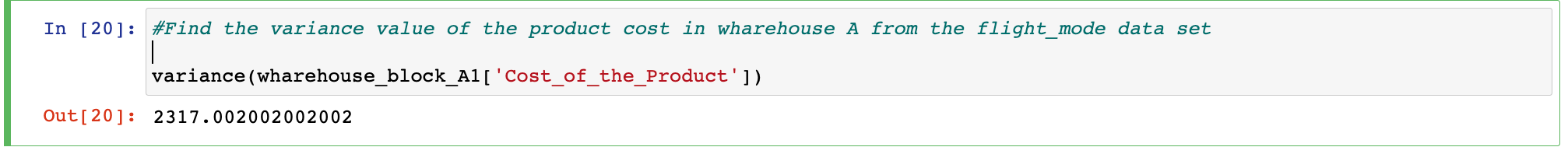
* Mode Value = 204



* **Measure of spread (variance, standard deviation) of Variable “Cost\_of\_the\_Product”**

~ For Shipment type = Flight

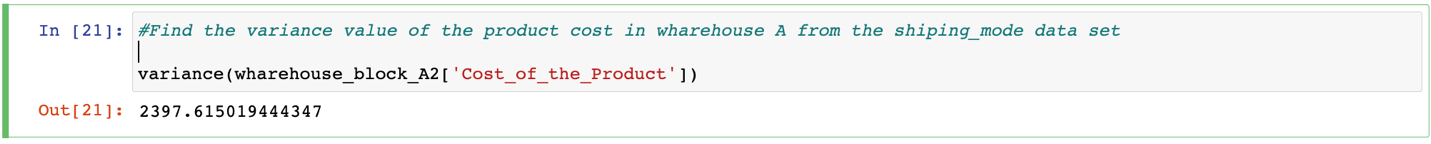
* Variance Value = 2371.002

****

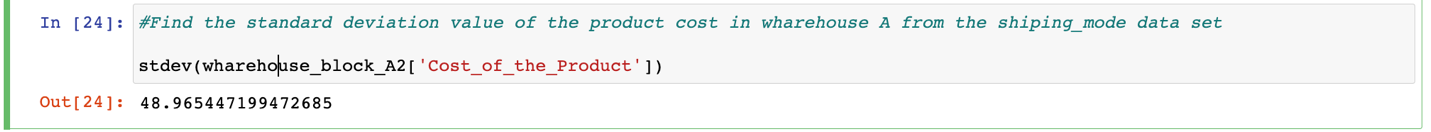
* ****Standard Deviation Value = 48.135

~ For Shipment type = Ship

* Variance Value = 2397.615

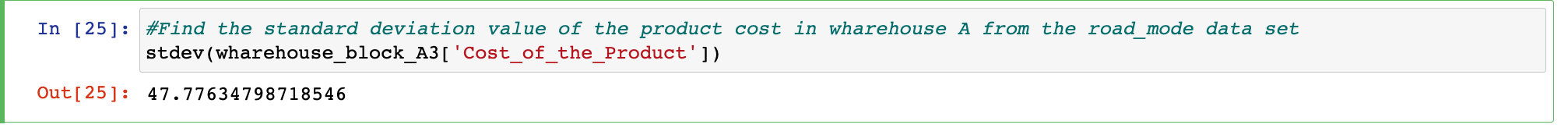


* Standard Deviation Value = 48.965



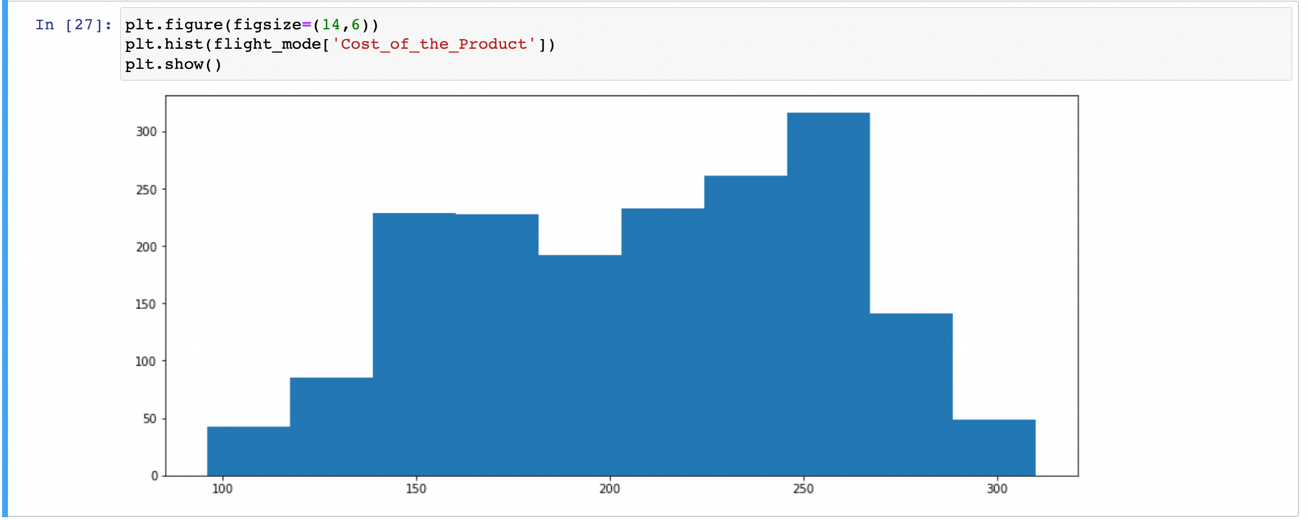
~ For Shipment type = Road

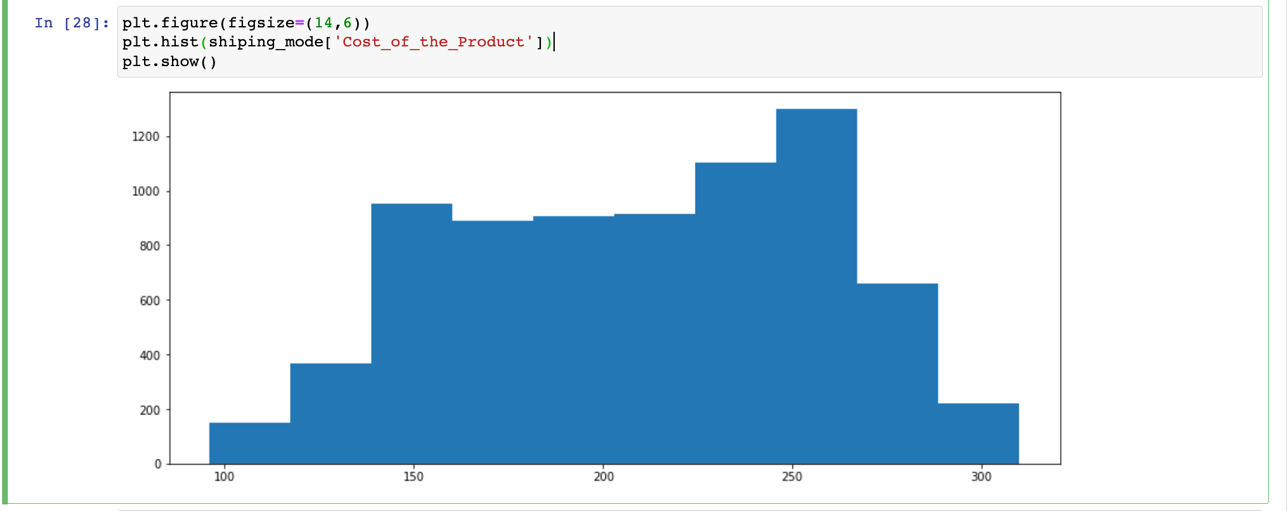
* Variance Value = 2282.579
* Standard Deviation Value = 47.776



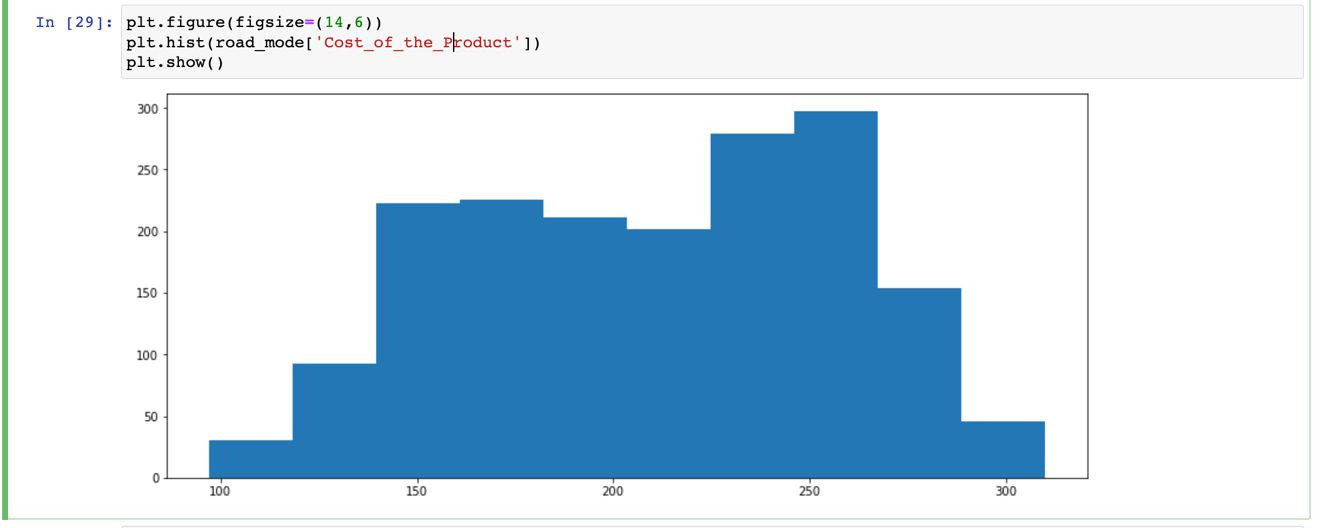
* Show the distribution of the “Cost\_of\_the\_Product” in histogram

~ For Shipment = Flight



~ For Shipment = Ship

~For Shipment = Road



* **Hypothesis testing (P-Value, T-Test, Chi-Square Test, ANOVA Test, Pearson Correlation, spearman rank correlation)**

**~** For Shipment = Flight

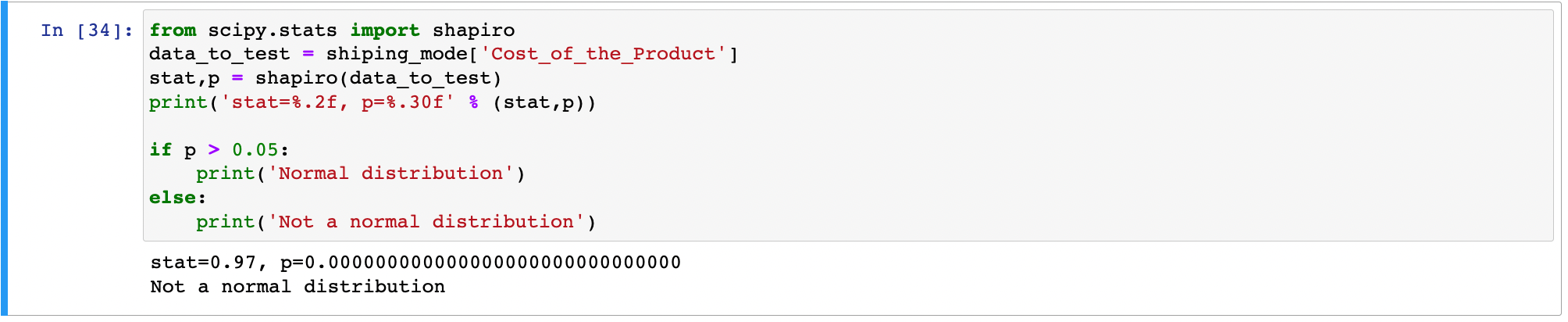
* Statistic = 0.97, P-Value = 0.000000000000000007409235417650

Not a normal distribution.



**~** For Shipment = Ship

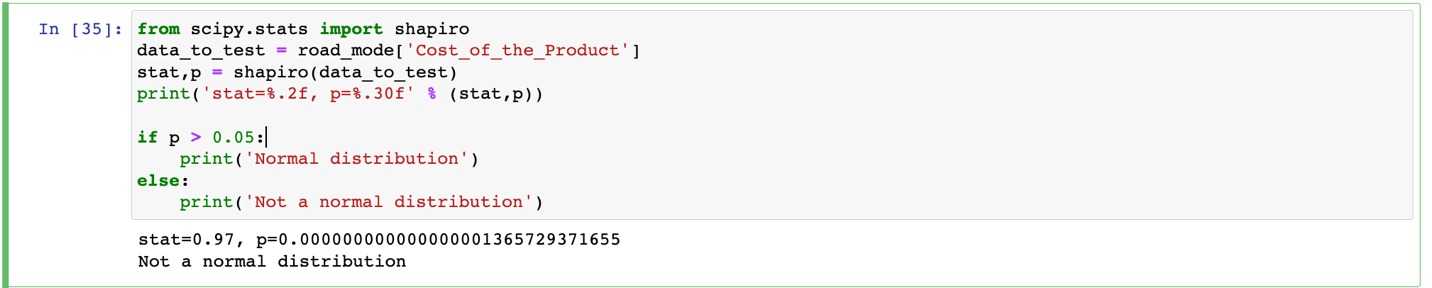
* Statistic = 0.97, P-Value = 0.00000000000000000

Not a normal distribution.

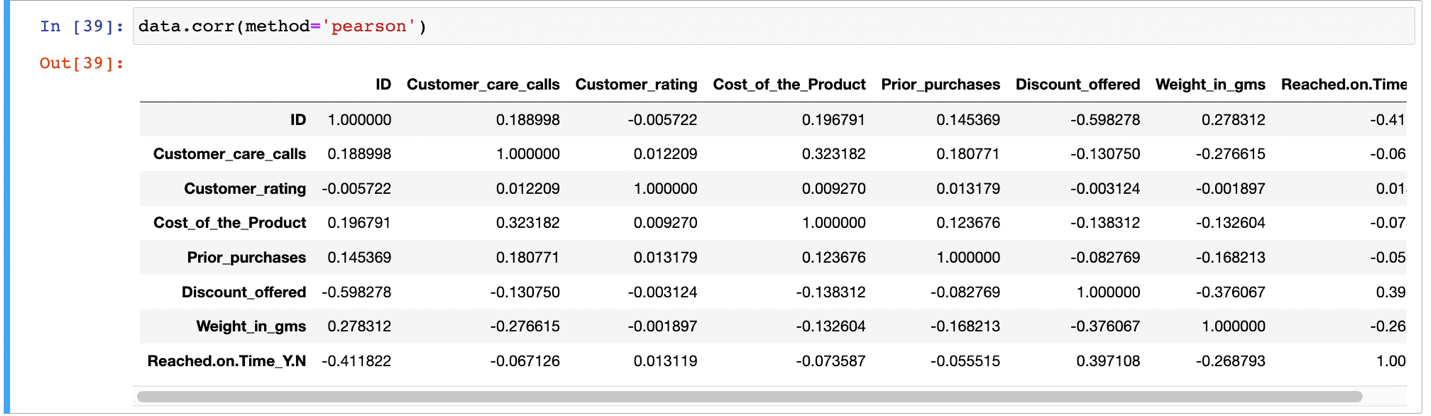
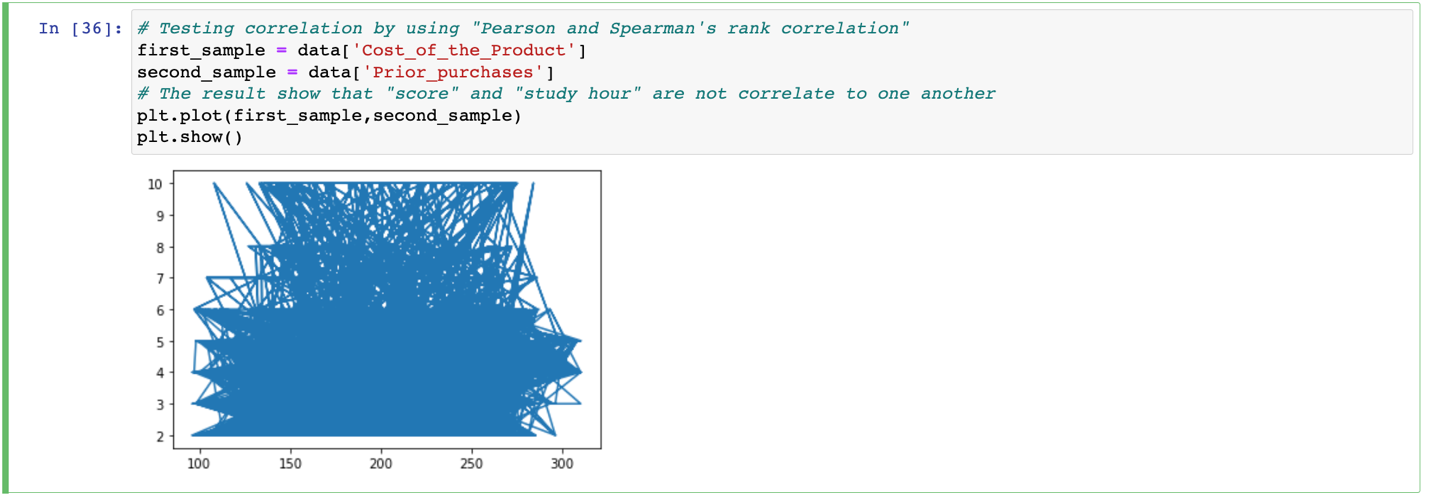
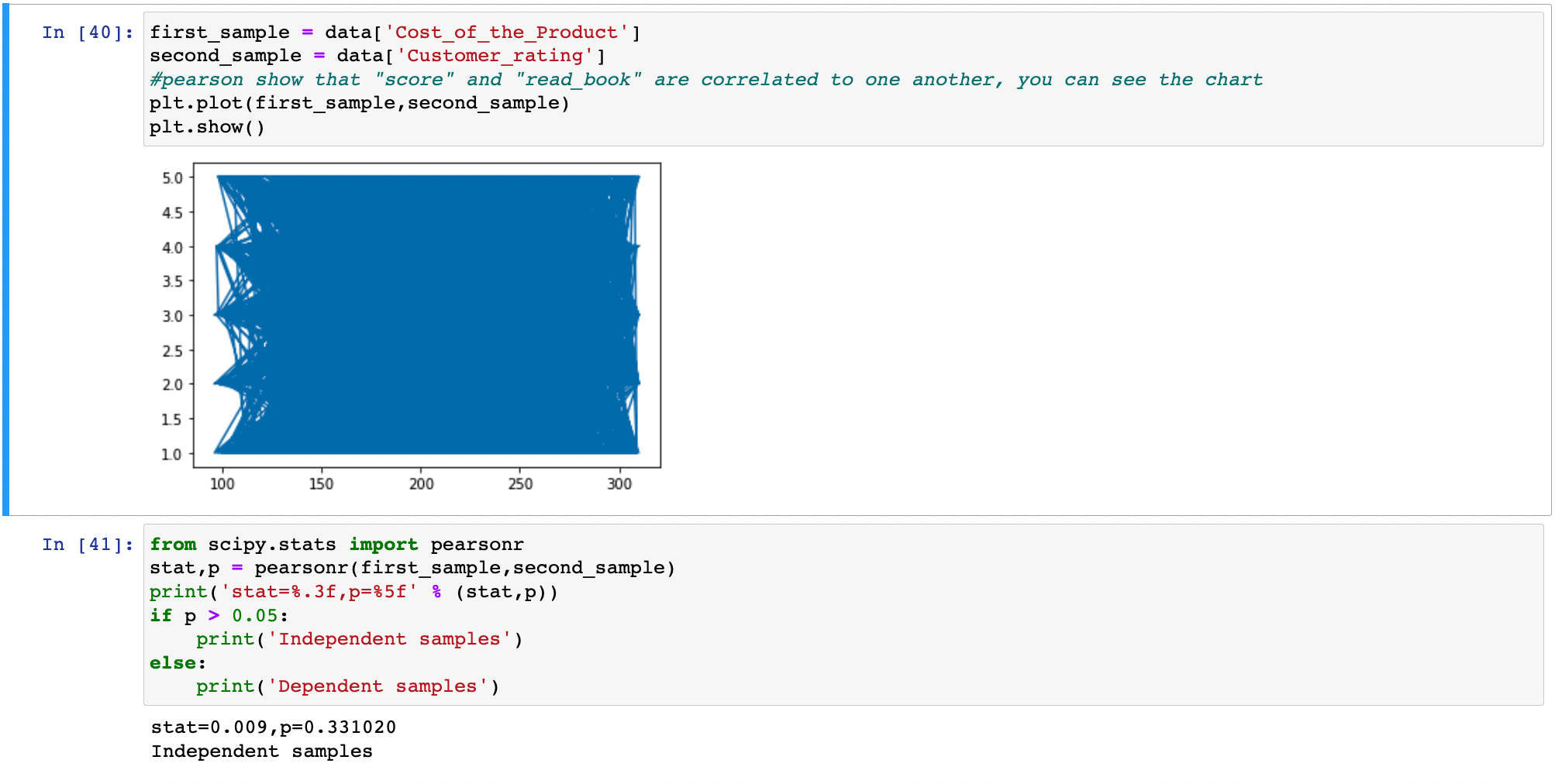
**~** For Shipment = Road

* Statistic = 0.97, P-Value = 0.000000000000000001365729371655

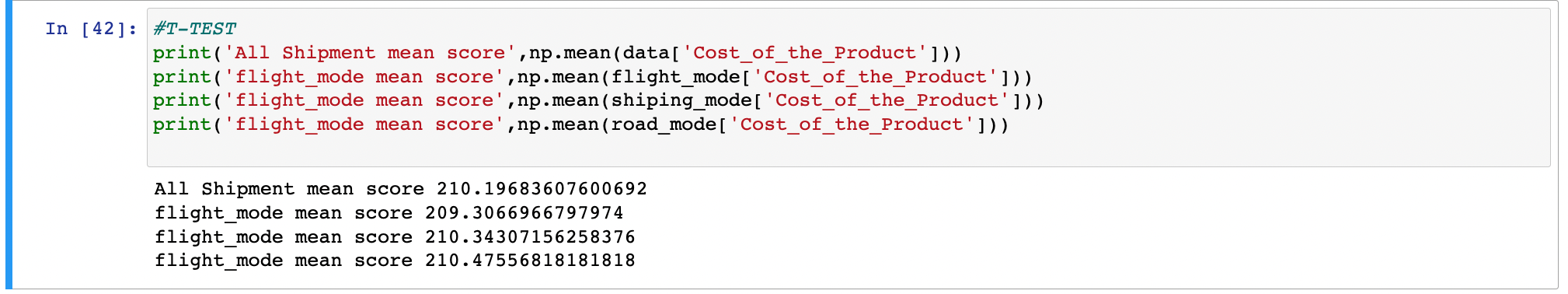
Not a normal distribution.



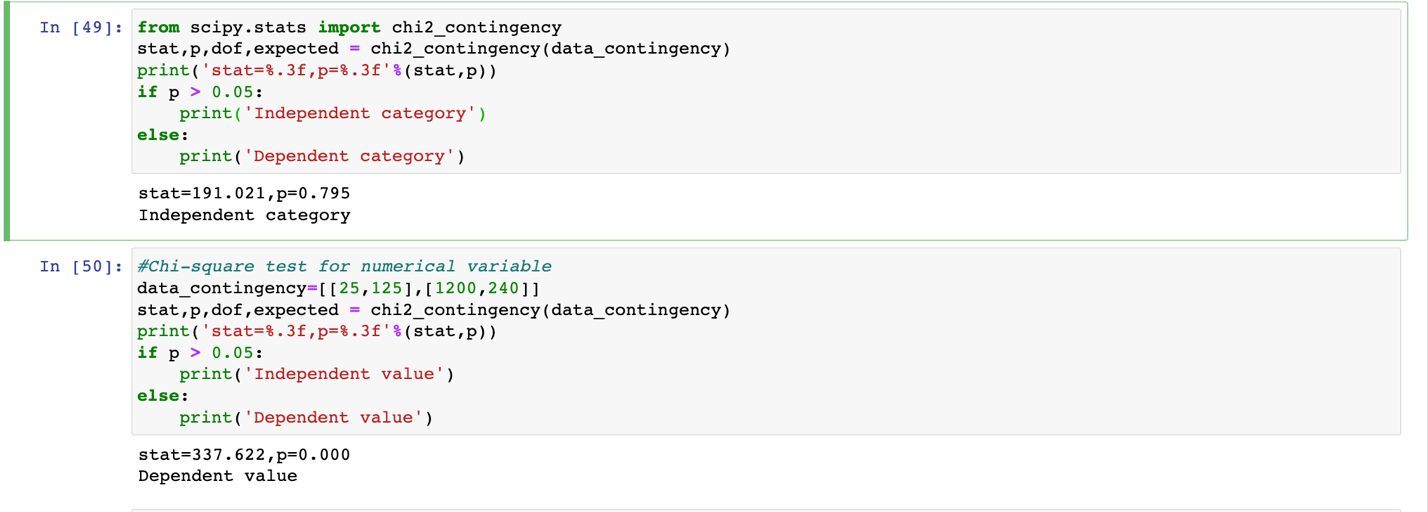
* + **Pearson Correlation**



* **T-Test**



* **Chi-Square**



* **Conclusion**

After doing some operation of statistic and go deeper into this dataset, we can see that mostly of the E-Commerce system used shipping type of shipment for delivery. Most customers are female who always order or buy product from the other countries.