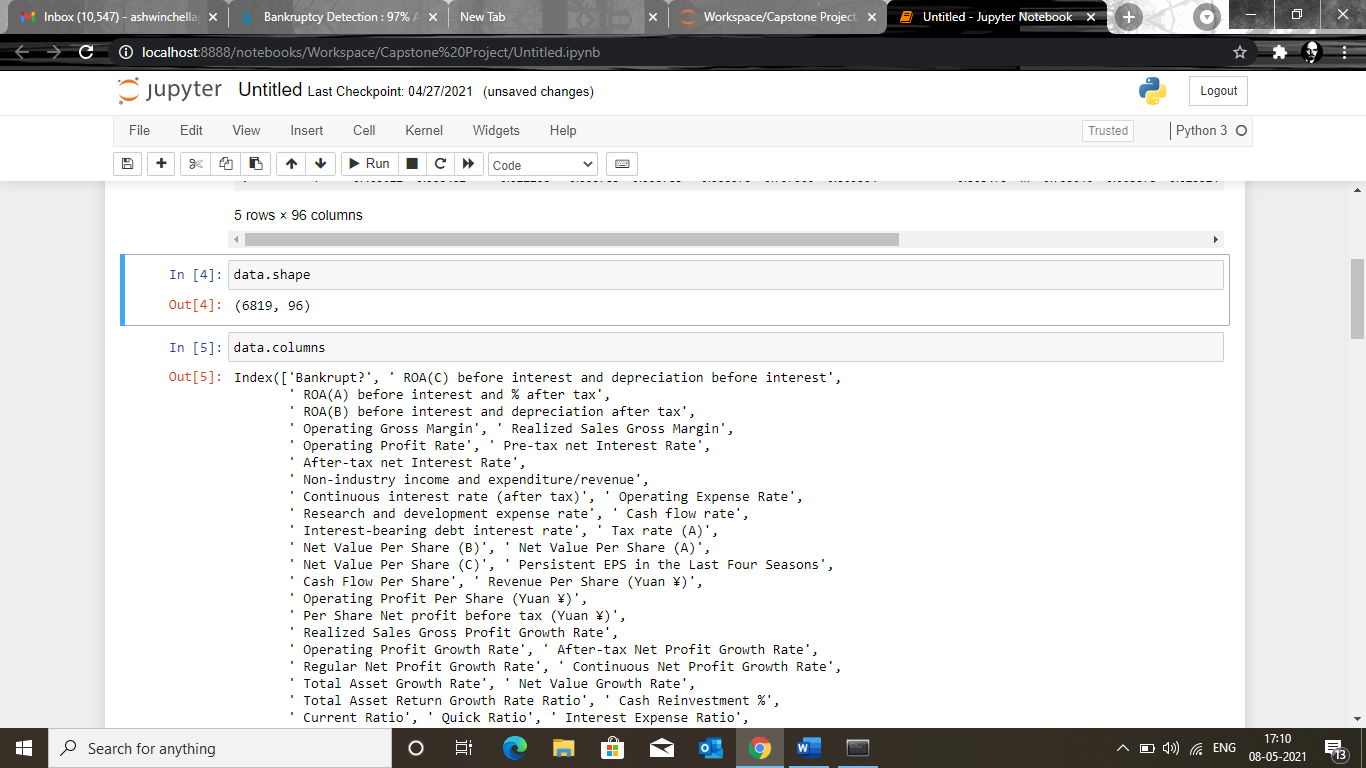
EDA Process

|  |  |  |
| --- | --- | --- |
| Si.No | TOPICS | Pg.No |
| 1 | * 1. **Shape of Dataset**   2. **Describe** | 2  2 |
| 2 | **2.1 Column Names****2.2 Check constant Columns****2.3 Check Duplicates****2.4 Check Missing Values** | 2  3  3  3 |
| 3 | **3.1 Target Class distribution****3.2 Outliers Handling****3.3 Observations** | 3  4  6 |

|  |  |
| --- | --- |
| **Author** | Ashwin C |
| **Date** | 08-05-2021 |
| **Release** | Version 1.0 |

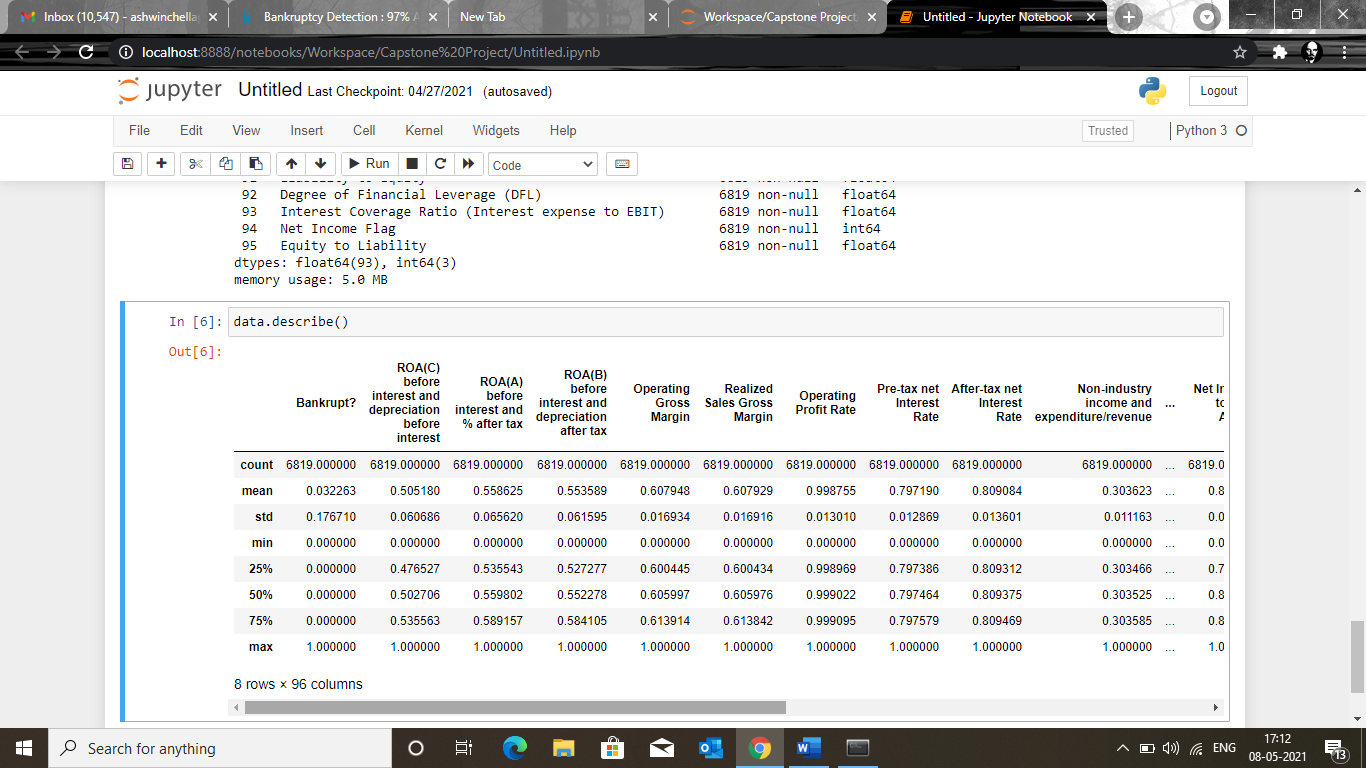
### **1.1 Shape of Dataset**

* We have total 6819 records from 96 columns in the dataset,of which:
  + 1 Target (Bankrupt?)
  + 95 features



### **1.2 Describe**

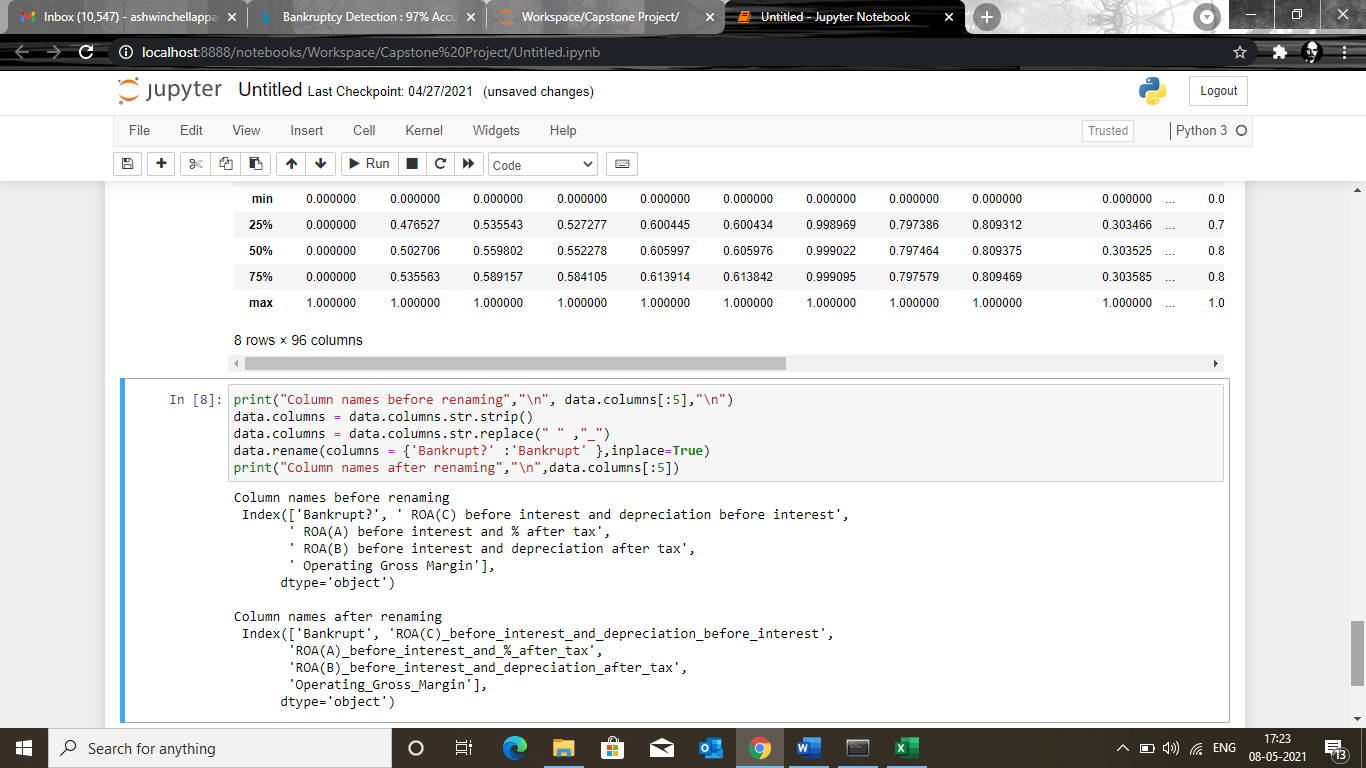
* All features are numeric
* one features has 0 variance that is constant throughout
* major features are in the range of 0-1
* there are outlier infected features



### **2. Sanity Check**

#### **2.1 Column Names:**

* Remove leading whitespaces from Column names
* Replace " " with "\_" in columnnames
* Rename Target column
* Rename **Bankrupt? -> Bankrupt**



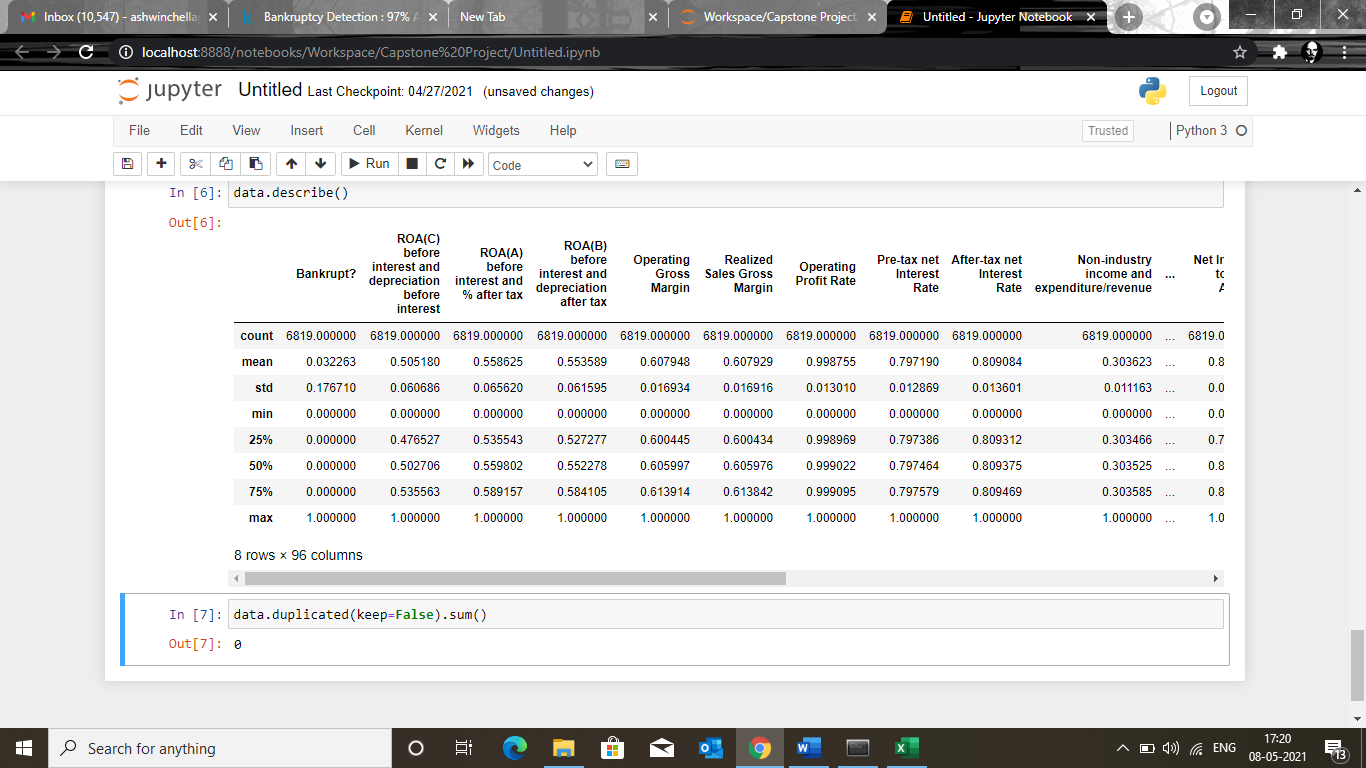
#### **2.2 Check constant Columns:**

* **Net Income Flag** is constant, hence drop it.

data.drop(['Net\_Income\_Flag'],axis=1,inplace=True) *## drop constant columns*

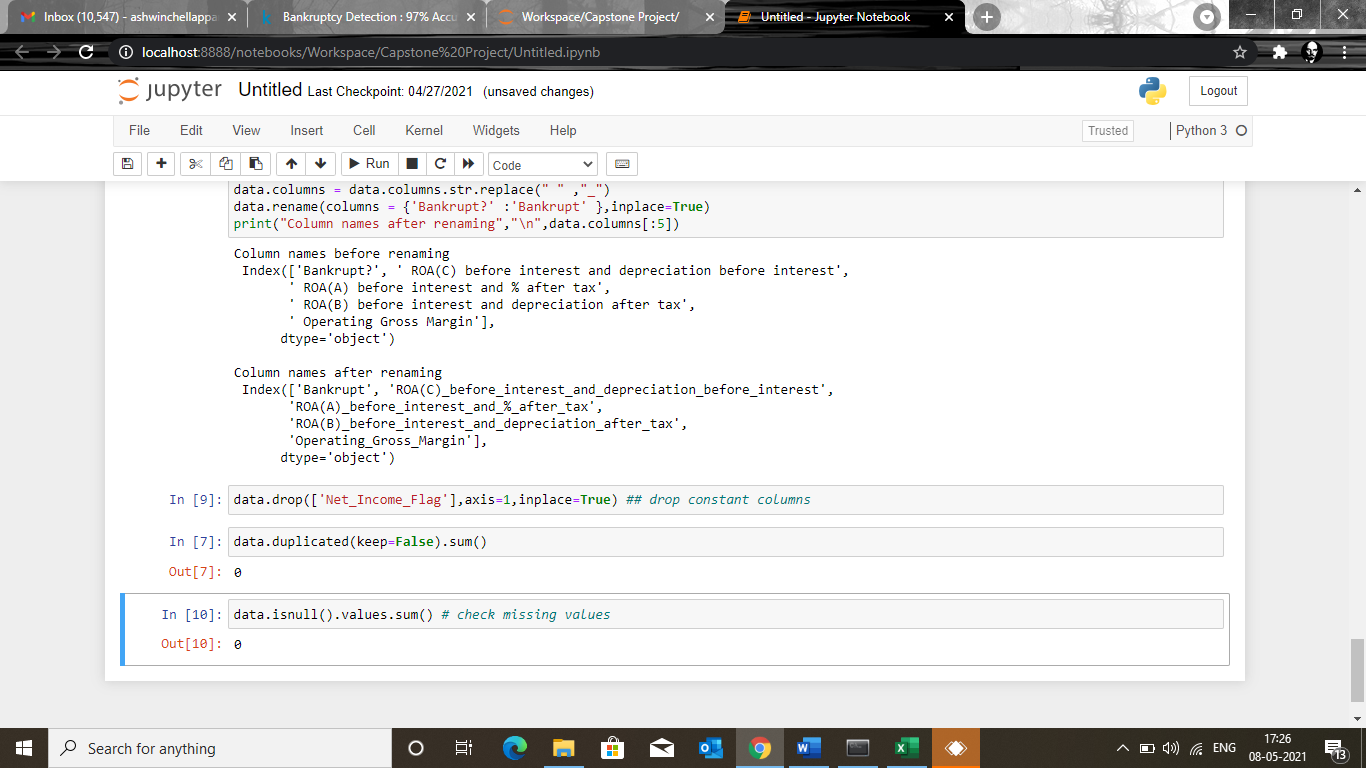
#### **2.3 Check Duplicates**

* No duplicated values



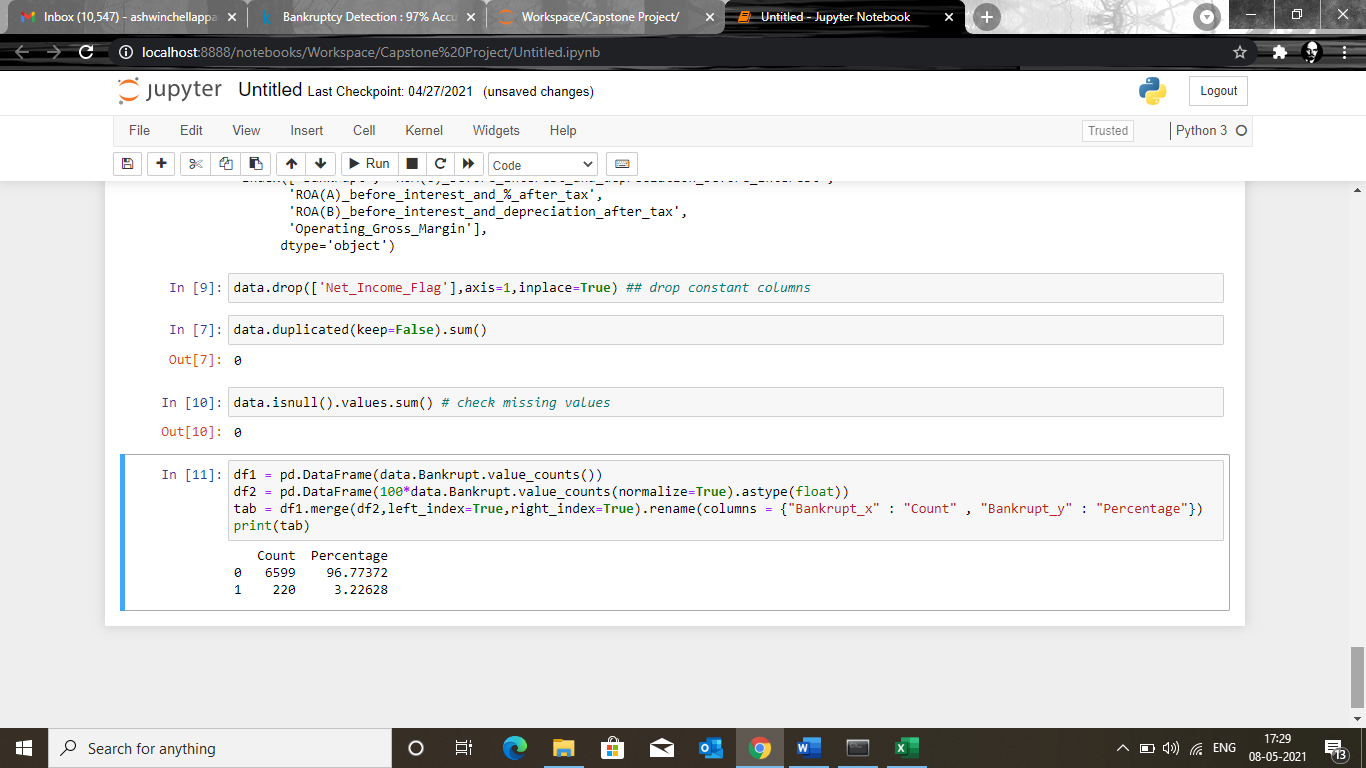
#### **2.4 Check Missing Values**

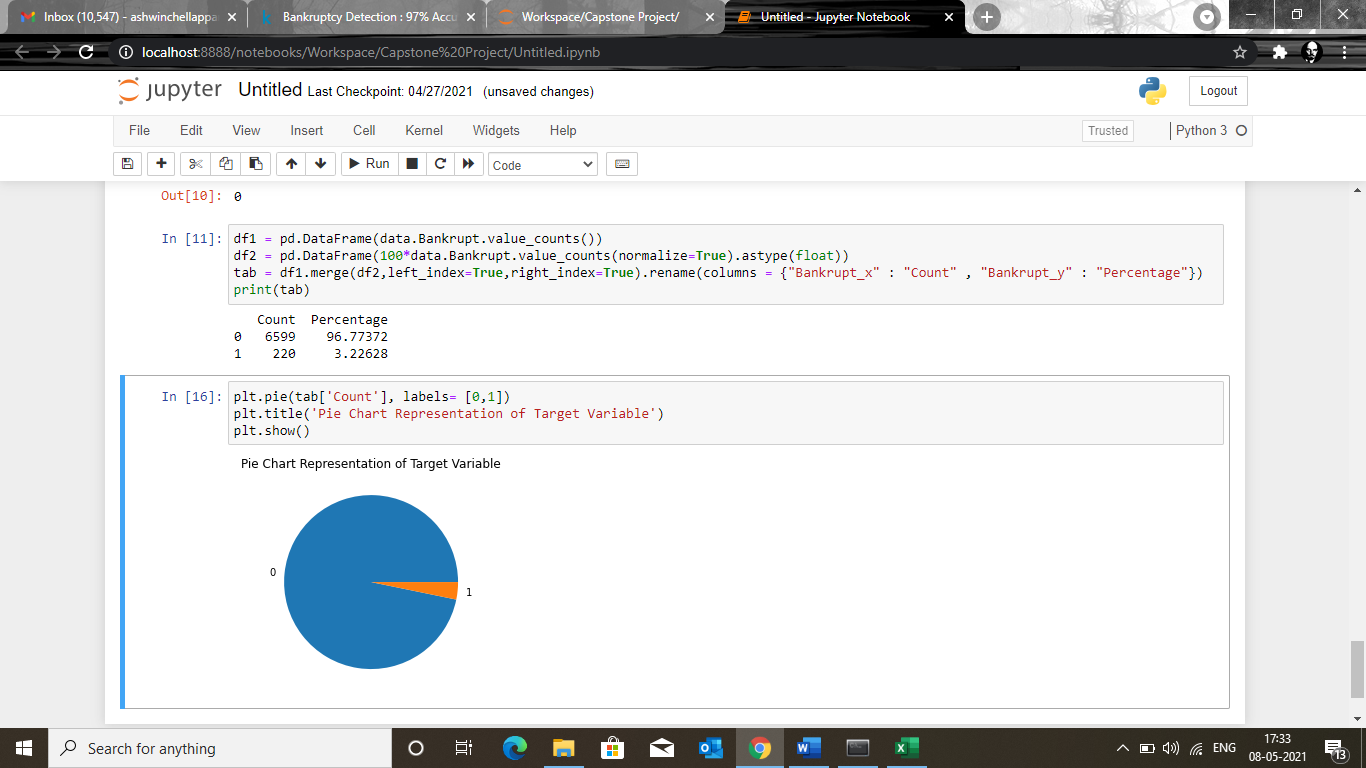
* No missing values



### **3.1 Target Class distribution**

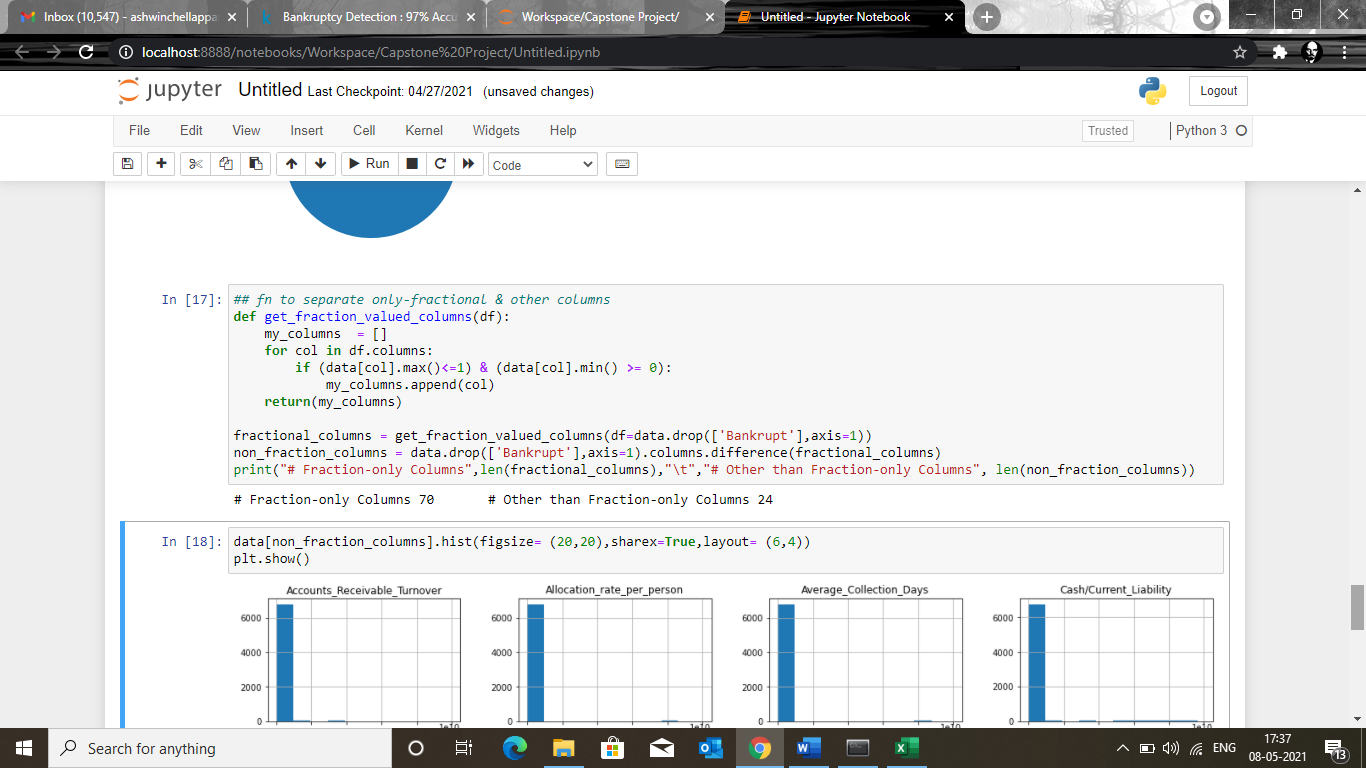
* Target is heavily imbalanced
* Bankruptcy Rate is around 3.2%

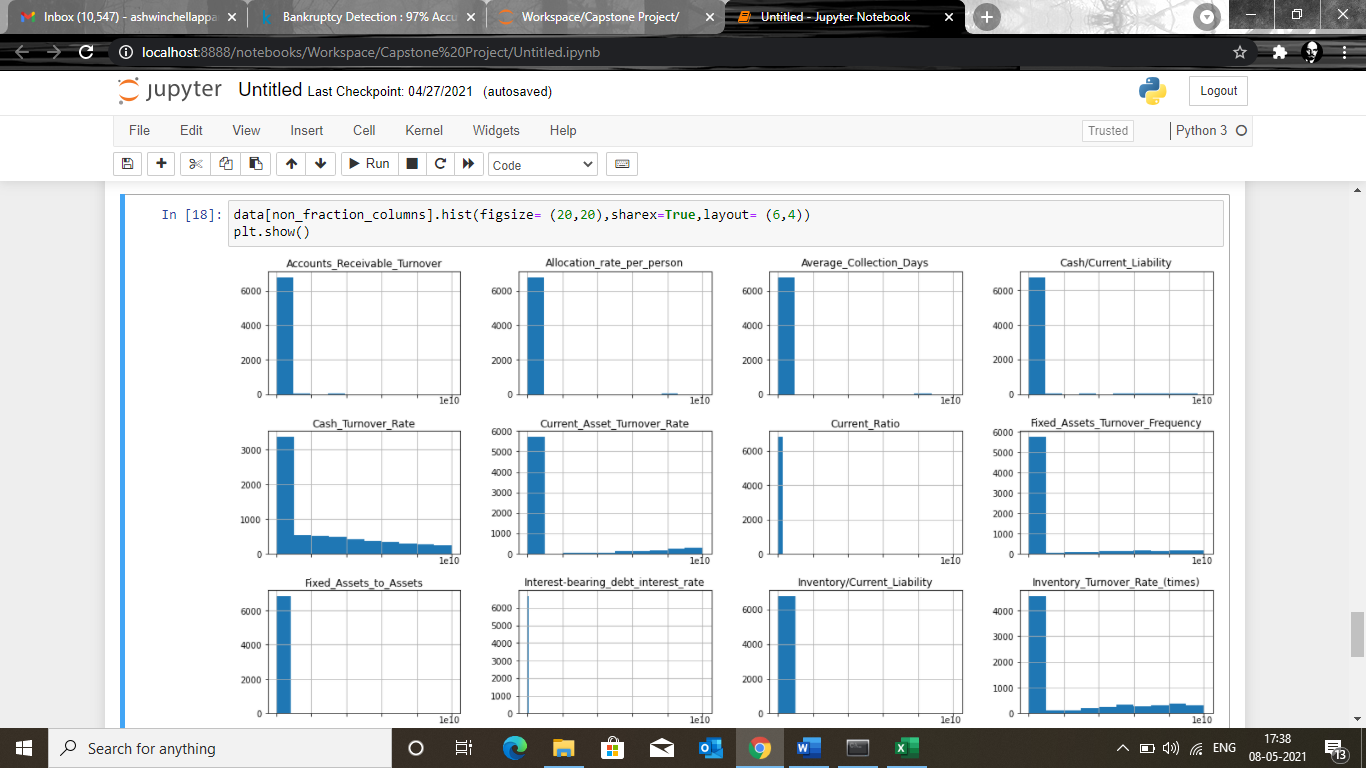


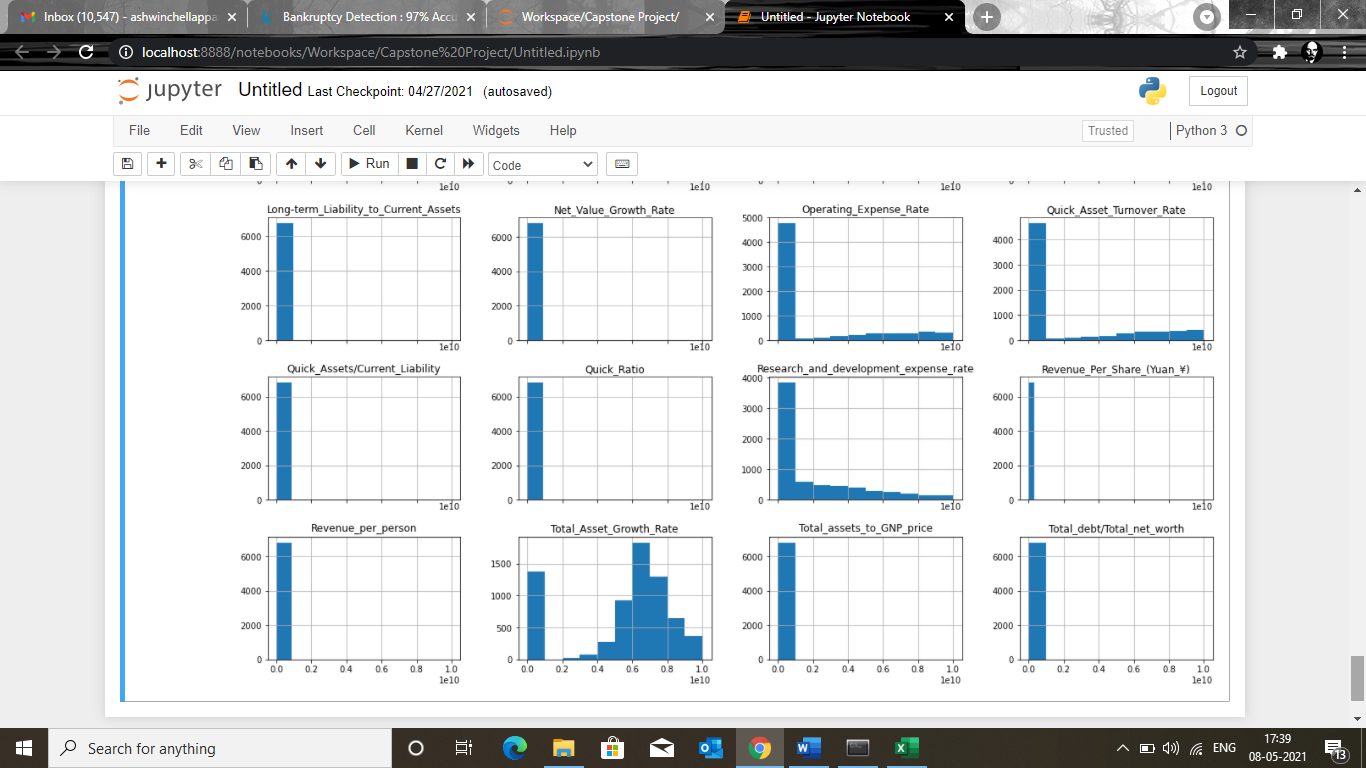


### **3.2 Outliers Handling**

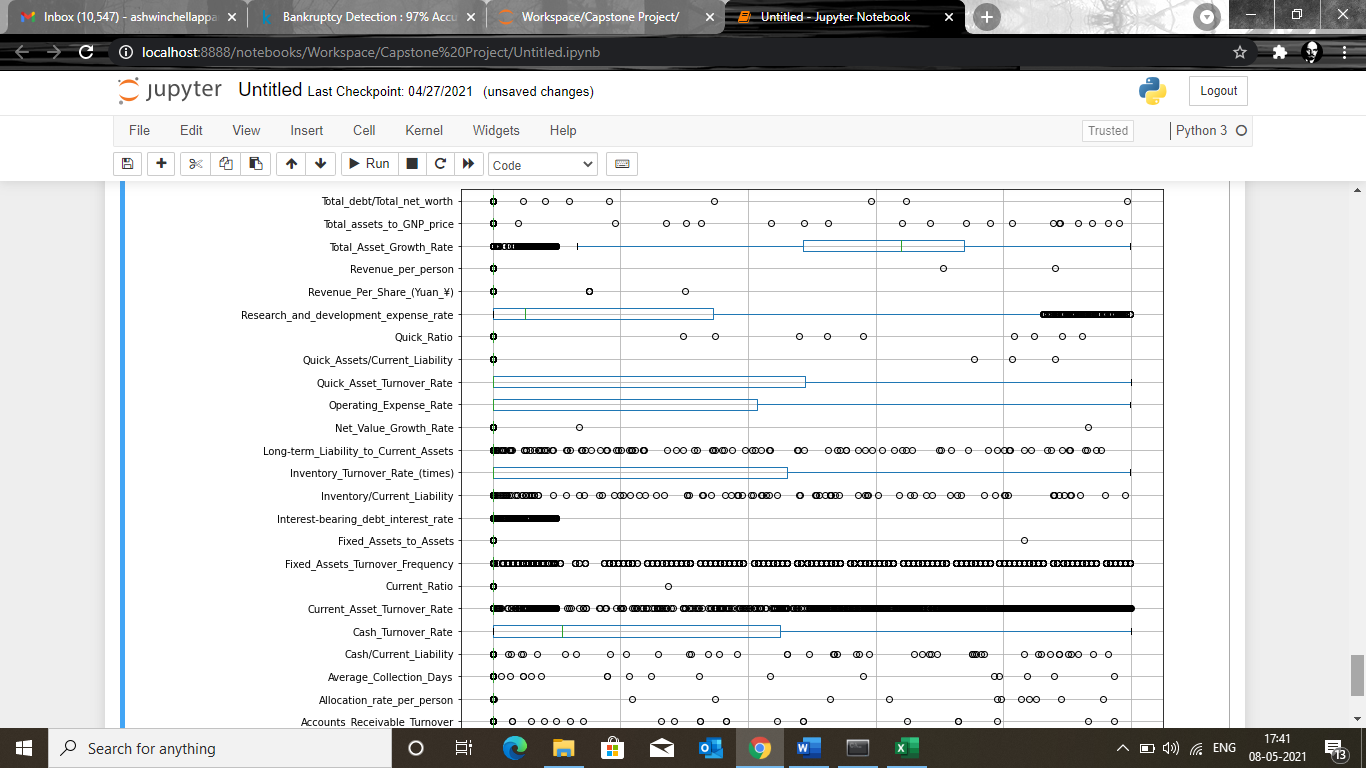
* First separate all **94** features into two groups
  + fraction-only features (i.e. features having values in [0,1])
  + other than fraction-only features
* **70** features are fraction-only features where as **24** are other than fraction-only features.
* Outliers are mainly **present** in these 24 "other than fraction-only" features
* To explore the outliers nature, distribution of these 24 features, are obtained using :
  + Histogram
  + Boxplot







### **Other than fraction-only features : Boxplot**



### **3.3 Observations**

* Major values are concentrated around starting ranges yet there are very high valued records.
* Some features show outliers in top 1% values only.Few of such features are:
  + Total\_debt/Total\_net\_worth
  + Revenue\_per\_person
  + Net\_Value\_Growth\_Rate
  + Revenue\_Per\_Share etc

1. There are some features that have significant number of higher values, like:
   * Current\_Asset\_Turnover\_Rate
   * Cash\_Turnover\_Rate

* **These features may potentially seem to be outliers but not be practically so.Hence outliers here need to be handled more cautiously,simply eliminating records lying above some cutoff cant be applied for all throughout**.
* We Need to have a rule:

#### **Rule :**

* Only **other than fraction-only** features are considered for outlier inspection
* Among them those meeting following conditions I decide to term them as outliers infected:
  + 100th percentile value is atleast 100 times greater than 99th percentile.
  + There are 10 or less records for 100th percentile.
* For the features satisfying above outlier condition we replace them as x-> log(1+x)
* **N.B: This is just a rule of thumb I made based on the experience while playing with the data.**

|  |
| --- |
| ['Accounts\_Receivable\_Turnover', 'Allocation\_rate\_per\_person', 'Average\_Collection\_Days', 'Cash/Current\_Liability', 'Current\_Ratio', 'Fixed\_Assets\_to\_Assets', 'Net\_Value\_Growth\_Rate', 'Quick\_Assets/Current\_Liability', 'Quick\_Ratio', 'Revenue\_Per\_Share\_(Yuan\_¥)', 'Revenue\_per\_person', 'Total\_assets\_to\_GNP\_price', 'Total\_debt/Total\_net\_worth'] |

|  |
| --- |
|  |