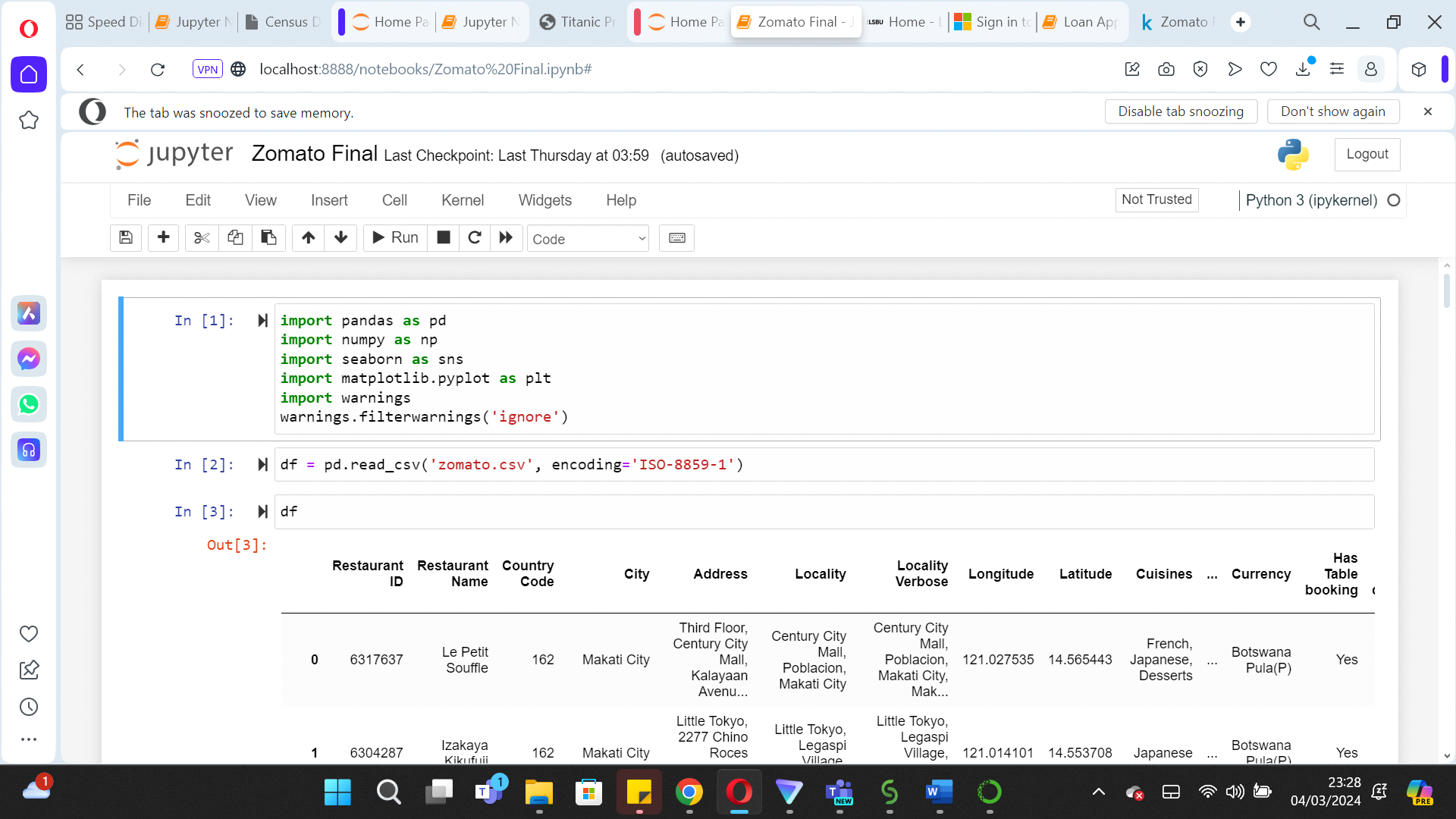
The Titanic dataset offers a compelling glimpse into the tragic event of 1912, blending historical insight with data analysis. The dataset is available on platforms such as Kaggle, it's a resource that draws not just data scientists, but anyone intrigued by the stories of those aboard the Titanic. This dataset contains the passengers’ demographics, including ages, genders, the number of siblings or spouses onboard, points of embarkation, and ultimately, their survival status. Key variables like Passenger ID, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked, provides a vivid picture of life aboard the Titanic, revealing the socio-economic layers that influenced survival chances.

PROBLEM STATEMENT

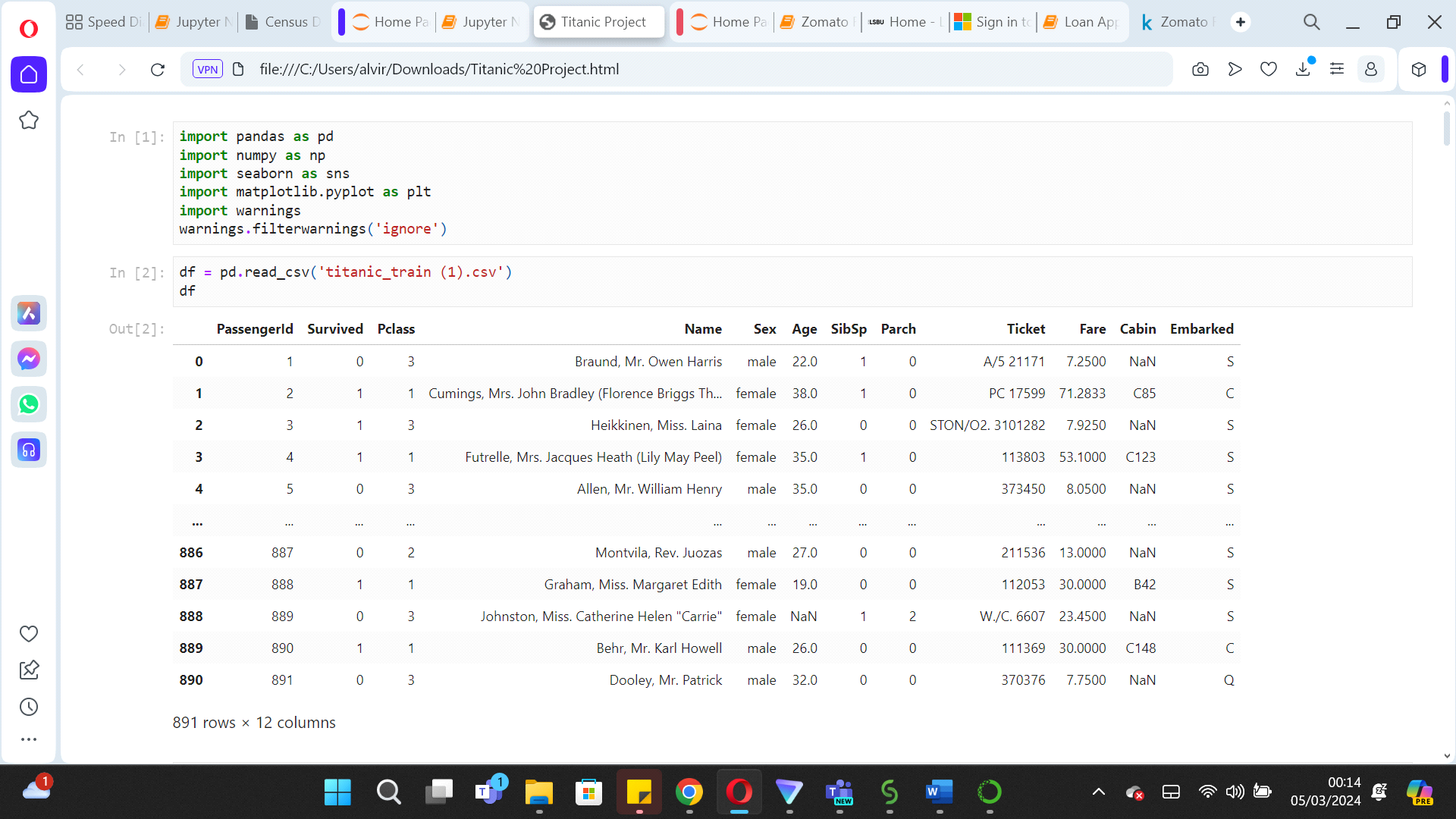
At its core, the Titanic dataset poses a question of survival, pushing us to explore the different factors that determine the survival of the passengers. It enables us to delve deeper and examine the roles of class, family connections, and even embarkation points in the survival narrative.

DATA ANALYSIS & EDA

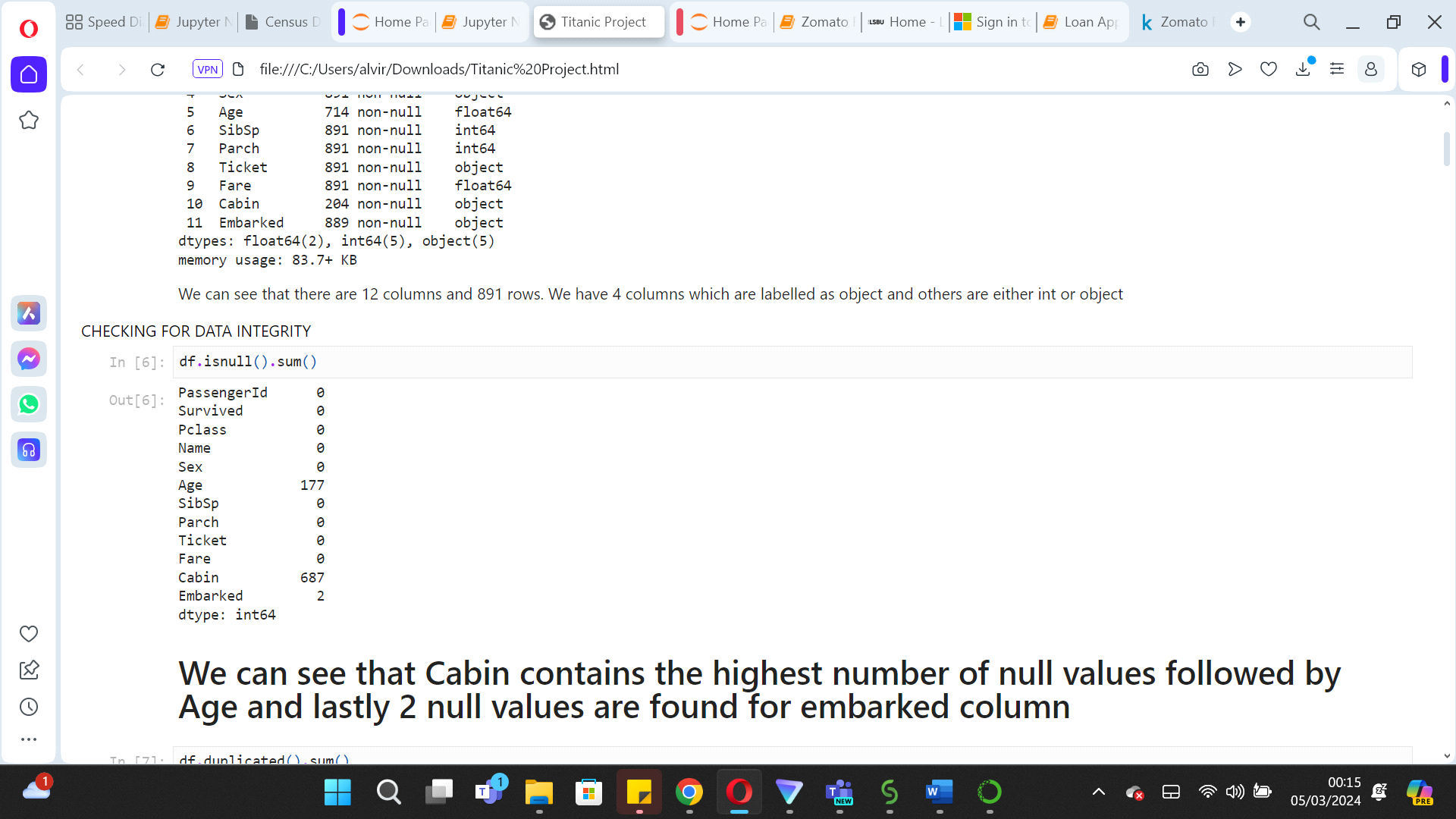
The very first step is to install the essential libraries & modules required to initiate the process of preparing and cleansing the data. The below code provides us with common computational code & visualisation libraries to understand our data in a better way.



The analysis then proceeds with importing the datasets, undertaking a preliminary examination to check its dimensions, structure, and integrity. This includes analysing the datasets for essential attributes such as size, shape, and detailed information, while identifying and sorting any anomalies present. These include missing values, duplicate records, or entries with irregularities such as question marks or extra whitespace. This step is quite crucial, as it ensures the datasets are thoroughly prepared and adjusted for the subsequent stages of in-depth analysis.





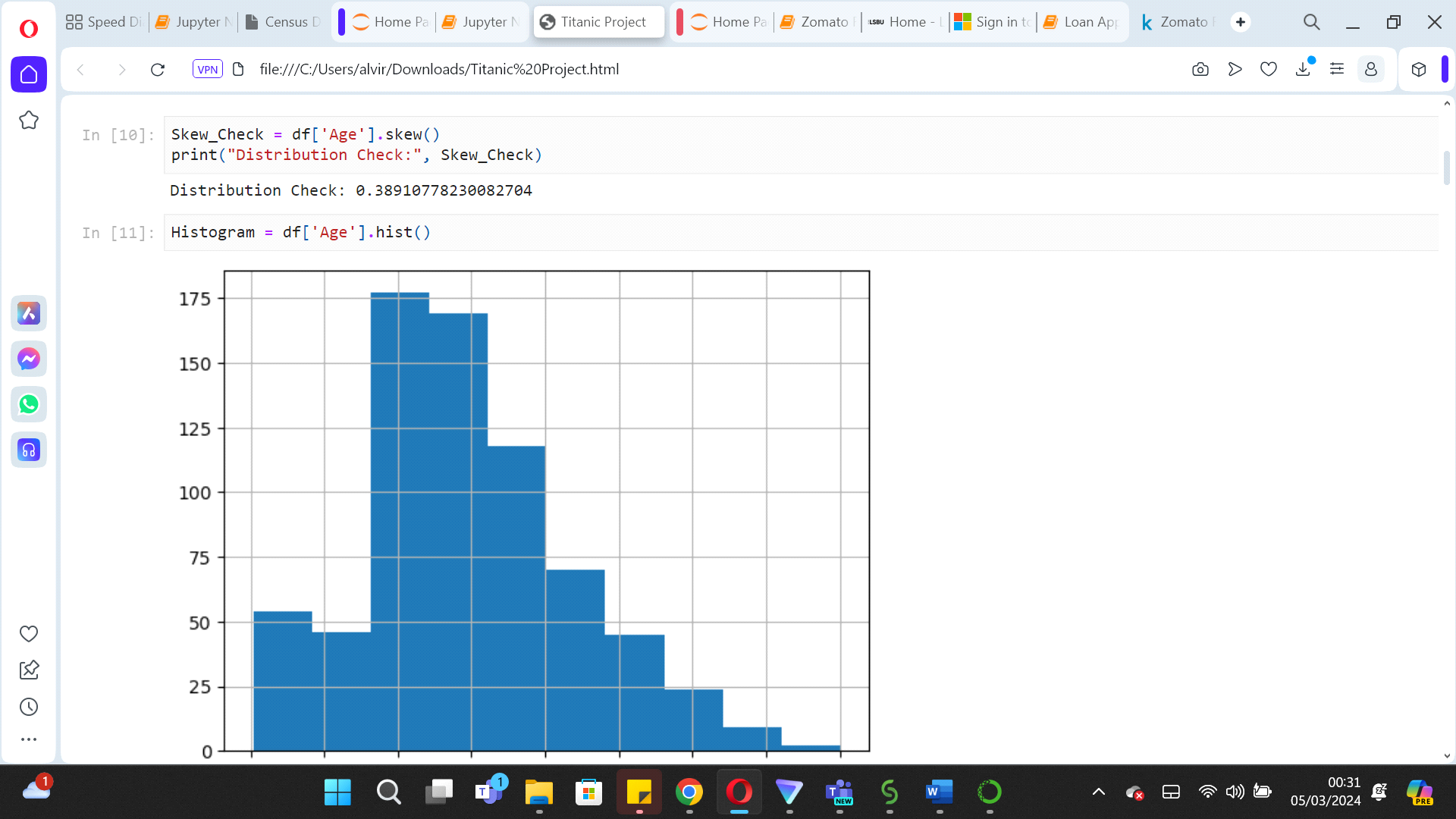


The Titanic dataset gives us a snapshot of the 891 passengers through 12 columns, which mixes text and numbers across different types of data. Out of these, 4 columns are text-based (like name, age & sex), and the rest are numbers, for example, age or fare. Moreover, the dataset also contain missing values for columns such as 'Cabin' 'Age' & ‘Embarked’.

Given that the 'Cabin' column is missing a significant portion of data, with 687 out of 891 values missing, it can be excluded from further analysis. The substantial number of missing entries compromises its utility especially in contributing any meaningful insights to our study.



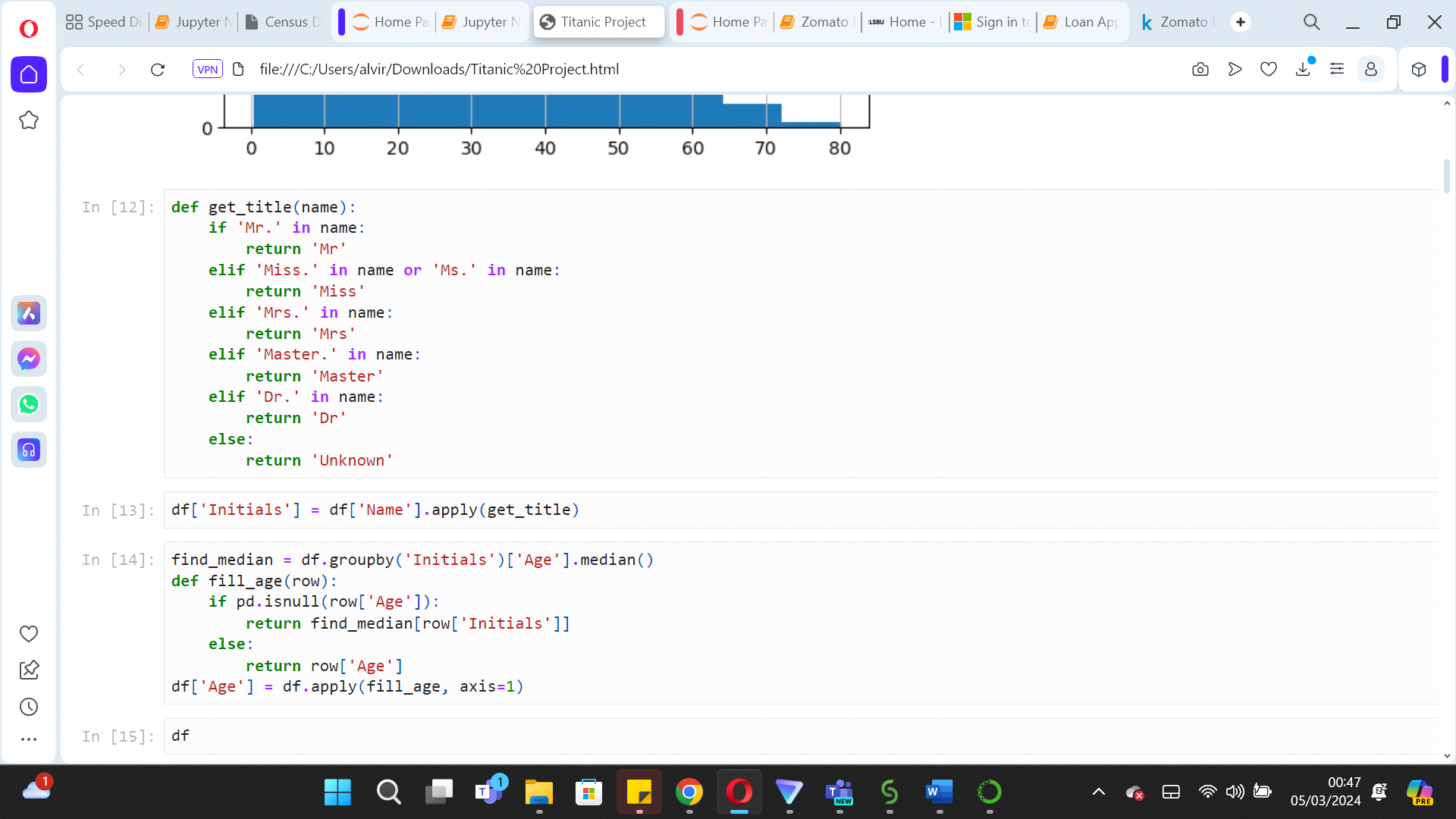
Continuing with the analysis, now focusing on the remaining 11 columns, it's crucial to address the null values in other columns thoughtfully. For instance, before deciding how to handle missing values in the 'Age' column, examining its distribution is important. This examination reveals that the 'Age' distribution is right-skewed, with a few outliers present. Such a distribution suggests that using the median to fill in missing values is more appropriate than the mean, as it is less influenced by outliers and provides a more accurate representation of the central tendency for age.



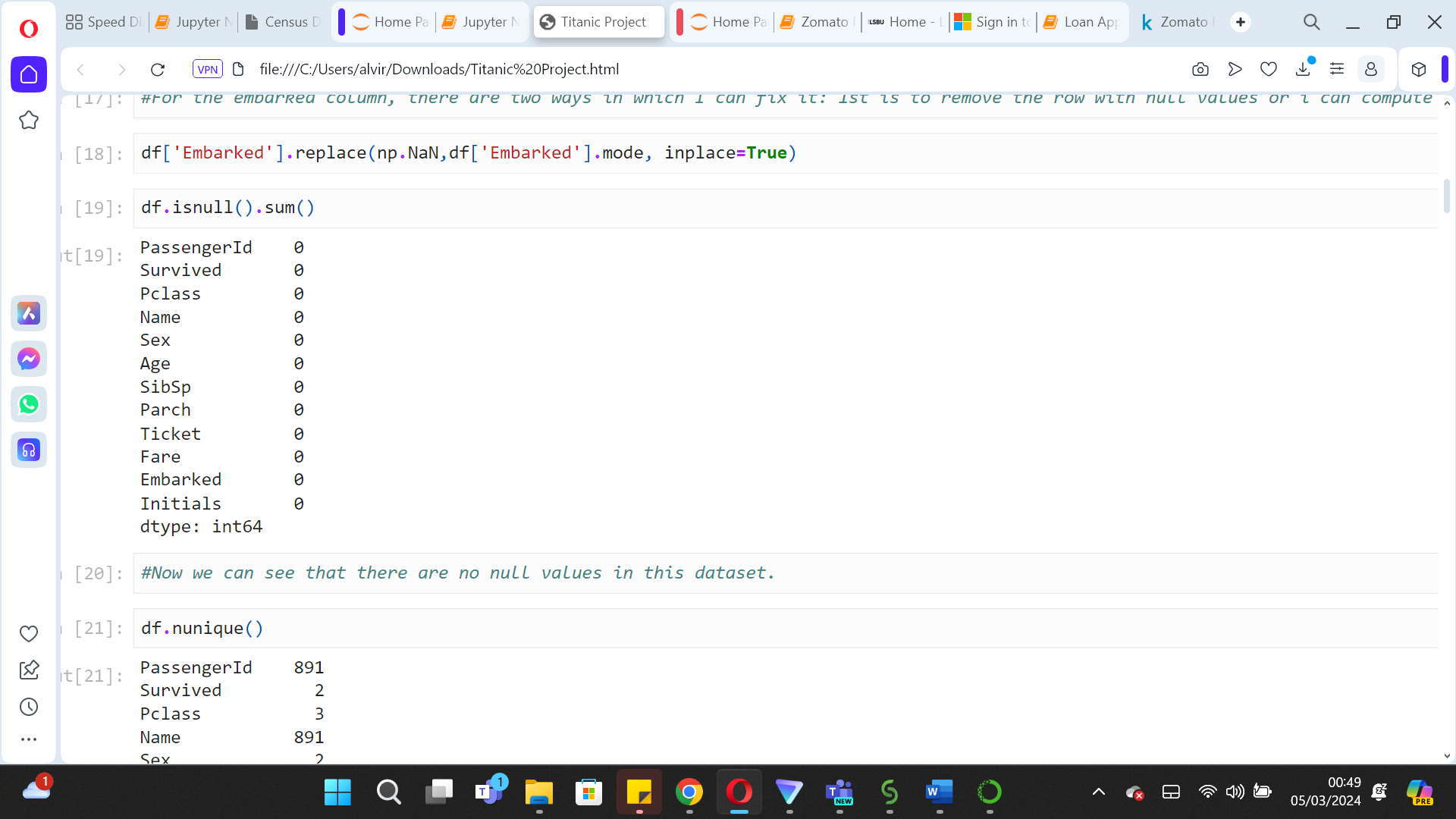
We can then employ a function as shown in the code below ‘get\_title’ to parse passenger names, extracting titles such as 'Mr', 'Miss', 'Mrs', and 'Dr'. These titles will provide us with clues about gender and marital status but could also imply age and social hierarchy. We can then apply this function to generate a new column labelled 'Initials', categorizing passengers accordingly.

Building on this categorization, we can then compute the median age for each title group using the groupby functionality on 'Initials'. It then addresses the missing values in the 'Age' column by assigning the median age relative to the extracted title, a step performed by the fill\_age function.

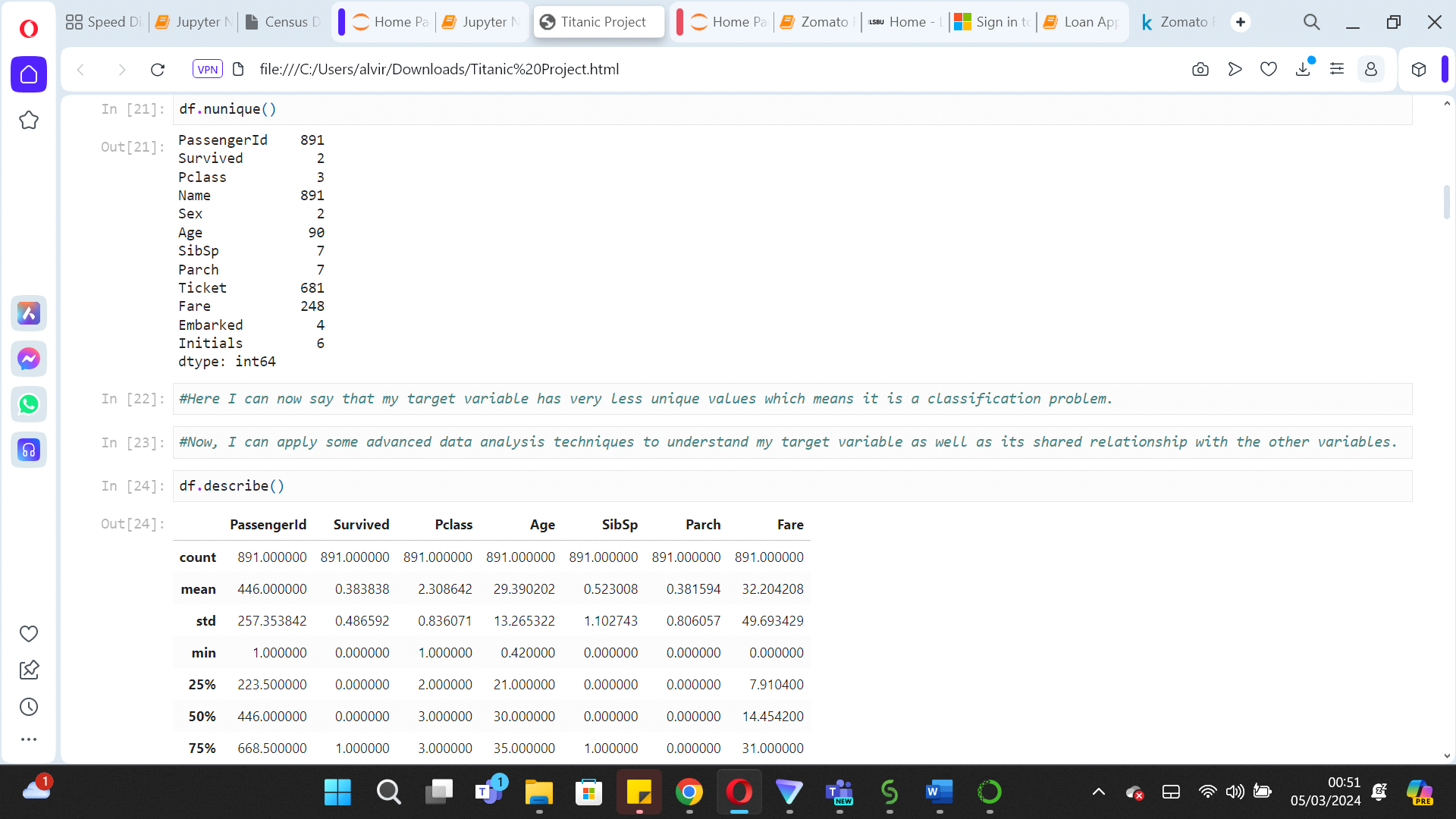
By implementing this method, the dataset gains improved accuracy in age data, which is important for any further analysis, particularly when predicting survival outcomes where age do tend to play a crucial role.



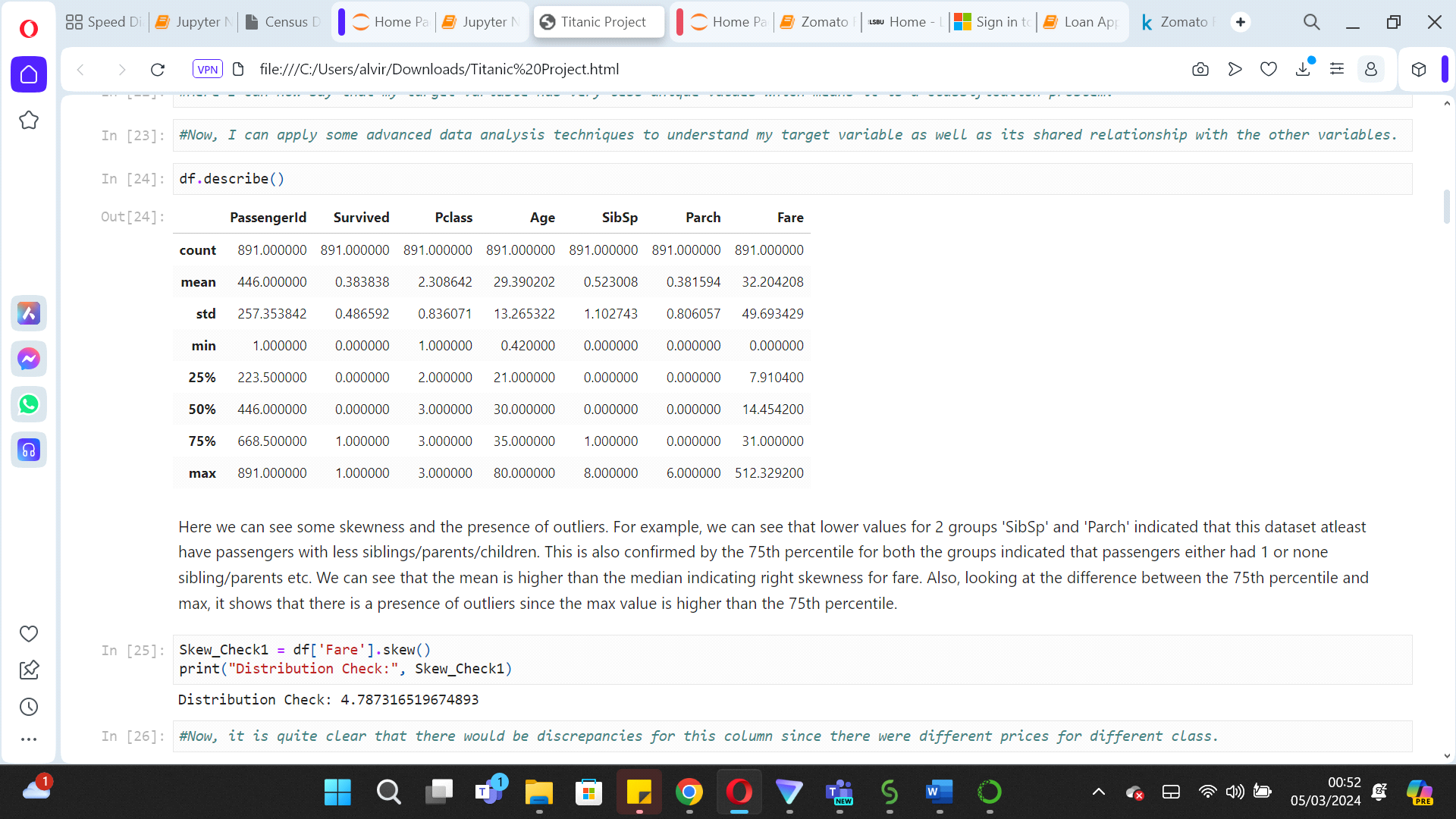
Moving further, when facing a minimal number of missing values, as is the case with the 'Embarked' column in the Titanic dataset which only has 2 missing entries, a common and straightforward approach is to fill in these gaps using the mode, which is the most frequently occurring value in the column. This is a very practical solution especially when the missing values are relatively insignificant and would not skew the overall data distribution.



The 'nunique' function in data analysis is particularly insightful because it does more than just count unique values; it aids in determining the type of problem we're dealing with. For our target variable 'Survived', the small number of unique values—just two, indicating 'survived' or 'did not survive'—confirms that we're tackling a classification problem, not a regression one. This distinction is crucial because it shapes the kind of predictive modelling approach we'll take, guiding us towards algorithms that are designed for categorical outcomes rather than continuous ones.



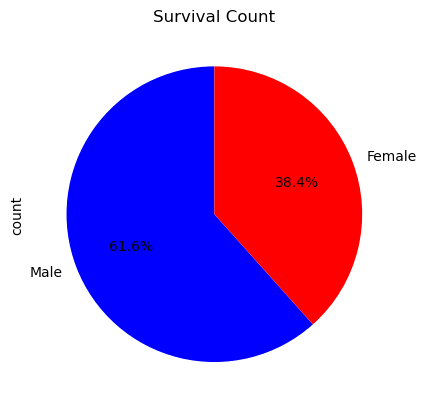
Delving into descriptive statistics furnishes us with deeper insights into the various variables, their distribution, central tendency, and the potential presence of outliers. For instance, a substantial gap between the 75th percentile and the maximum value can signal outliers, a trend we observe in columns like 'PassengerId', 'SibSp', 'Parch', and 'Fare'. Additionally, we encounter variables with skewed distributions; 'SibSp', 'Fare', and 'Parch' show evidence of right skewness with their means surpassing the medians, indicating a tail of higher values. Conversely, 'Survived', 'Pclass', and 'Age' are characterized by higher medians than means, suggesting left skewness. It is also noted that the average age hovers around 29, while the median age is approximately 32, further illustrating the dataset's varied distribution characteristics.



EXPLORATORY DATA ANALYSIS

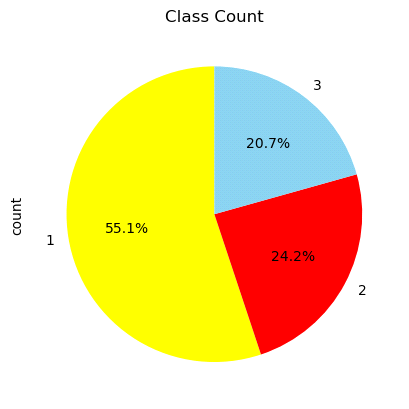
Exploratory Data Analysis (EDA) is a key step in data analysis, where data scientist conducts preliminary investigations to find patterns, identify any anomalies etc. This process is facilitated by employing both statistics and visual representations, providing an initial understanding that informs further analytics.

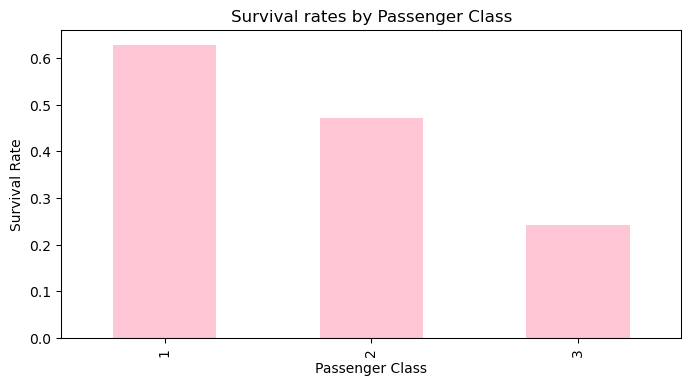
Analysing the target variable, it appears that there's an imbalance, as the number of individuals who did not survive significantly outweighs those who did.



This will need to be fixed in the later stages after the preliminary analysis. Further analysis can be done on various variables to see its impact on ‘Survived’ for example,

* Survival rates by passenger class



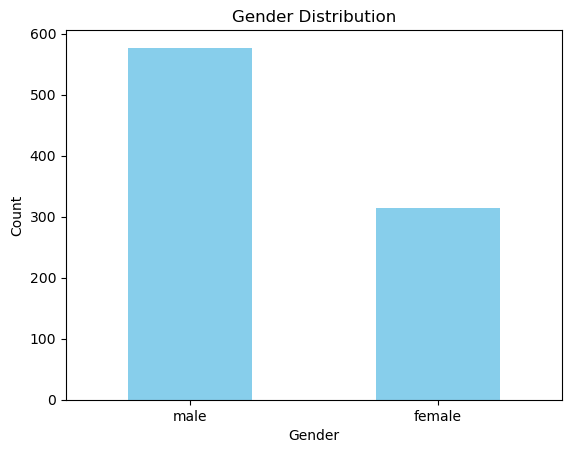


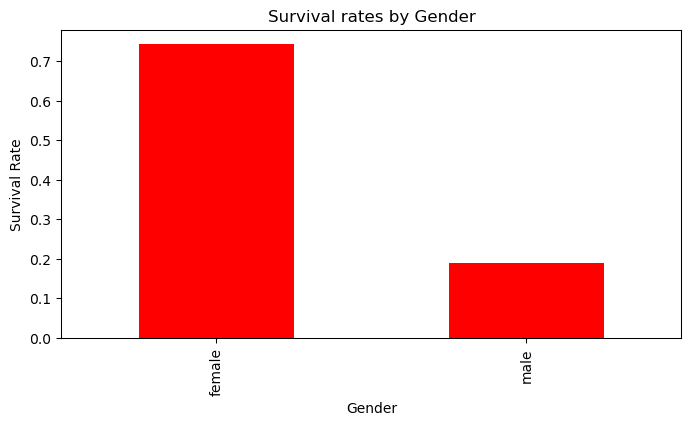
A graph of a number of people

Description automatically generated

It's evident that a significant number of passengers were in the 3rd class (491), outnumbering those in the second and first classes. Upon examining the survival rates, it becomes clear that 1st class passengers had markedly higher survival rates than those in the 2nd and 3rd classes.

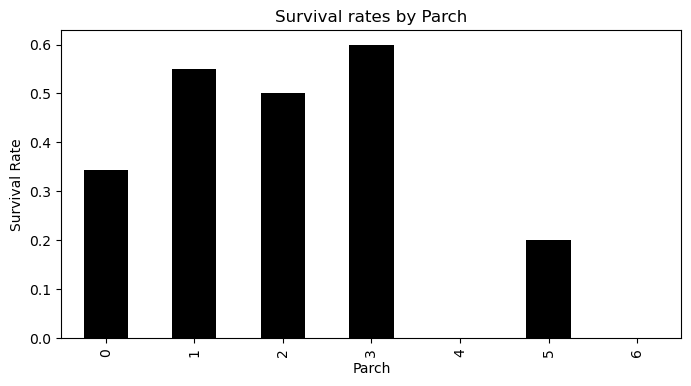
* Survival rates by Gender





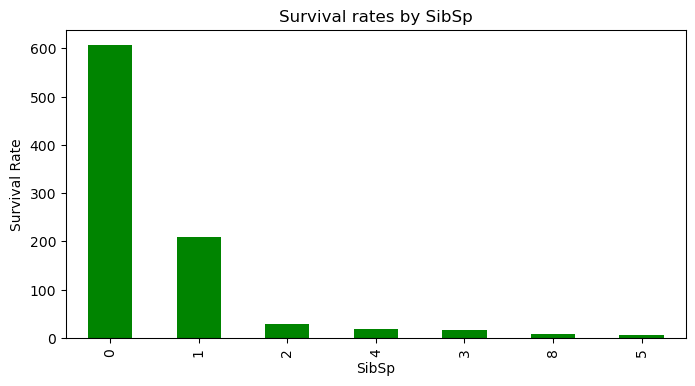
Despite the fact that there were more male passengers than female, the survival rate was notably higher for females. This observation lends credibility to the "women and children first" policy, suggesting it was indeed practiced during the evacuation.

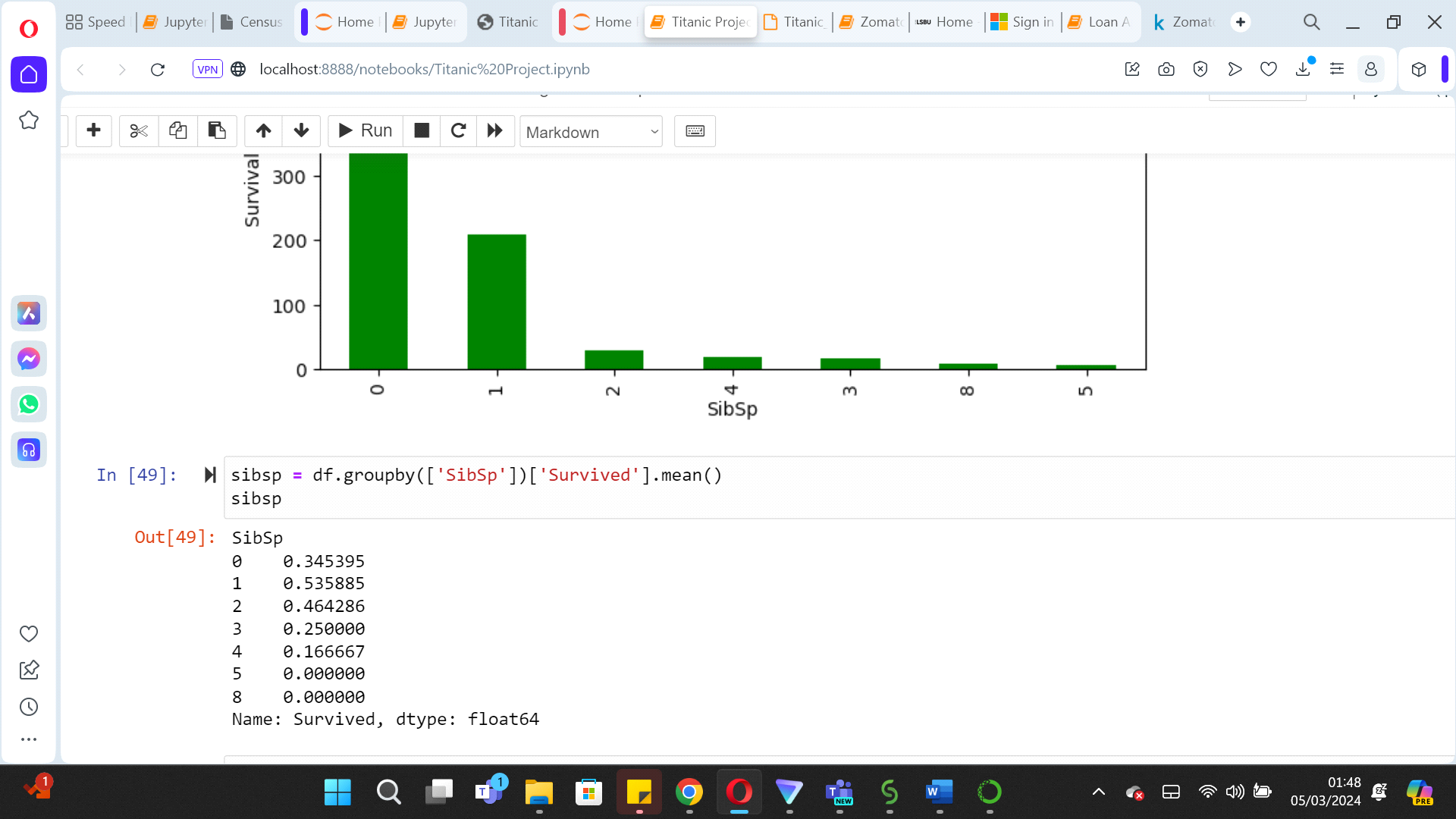
* Survival rates by Parch.



Individuals traveling alone had a lower survival rate compared to those accompanied by up to three parents or children, whose survival rates were notably higher. In fact, the group with three family members onboard shows a survival rate of 60%, ranking highest among the categories, followed by those with one, two, and then five family members.

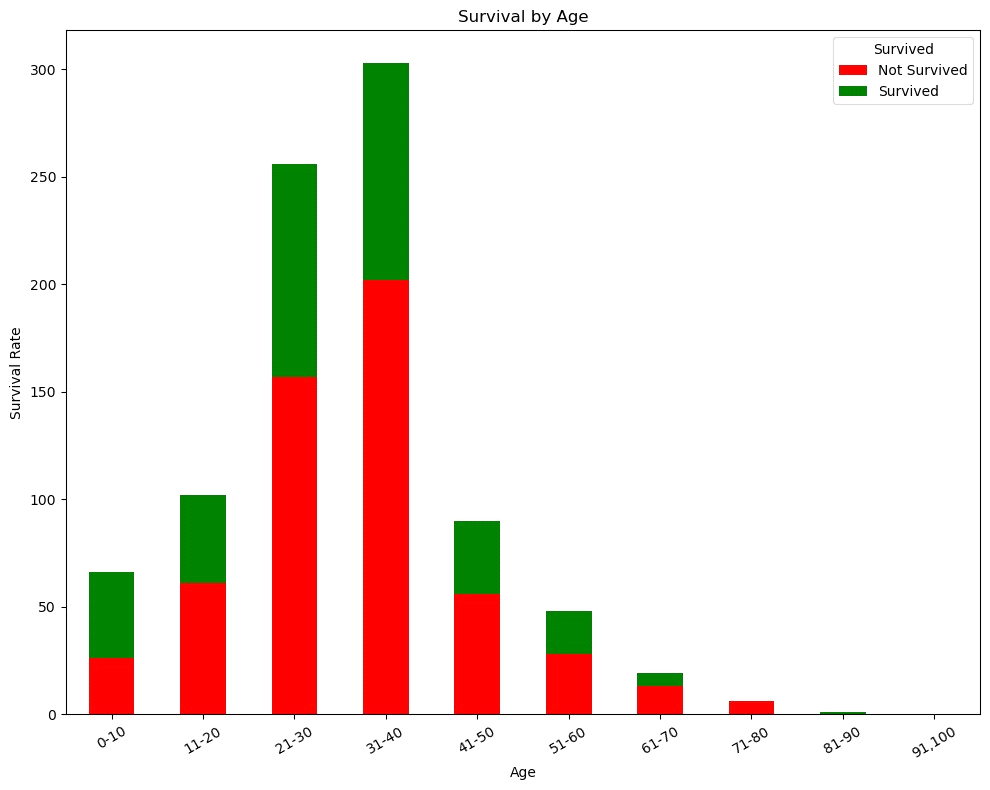
* Survival rates by SibSp





The graph illustrates that individuals with no siblings or spouses onboard (0 SibSp) initially appear to have a higher survival rate. However, when analyzing the data along with survival status, the mean values reveal an interesting pattern: passengers with at least one sibling or spouse (SibSp) present actually exhibit a greater likelihood of survival.

* Survival by Age

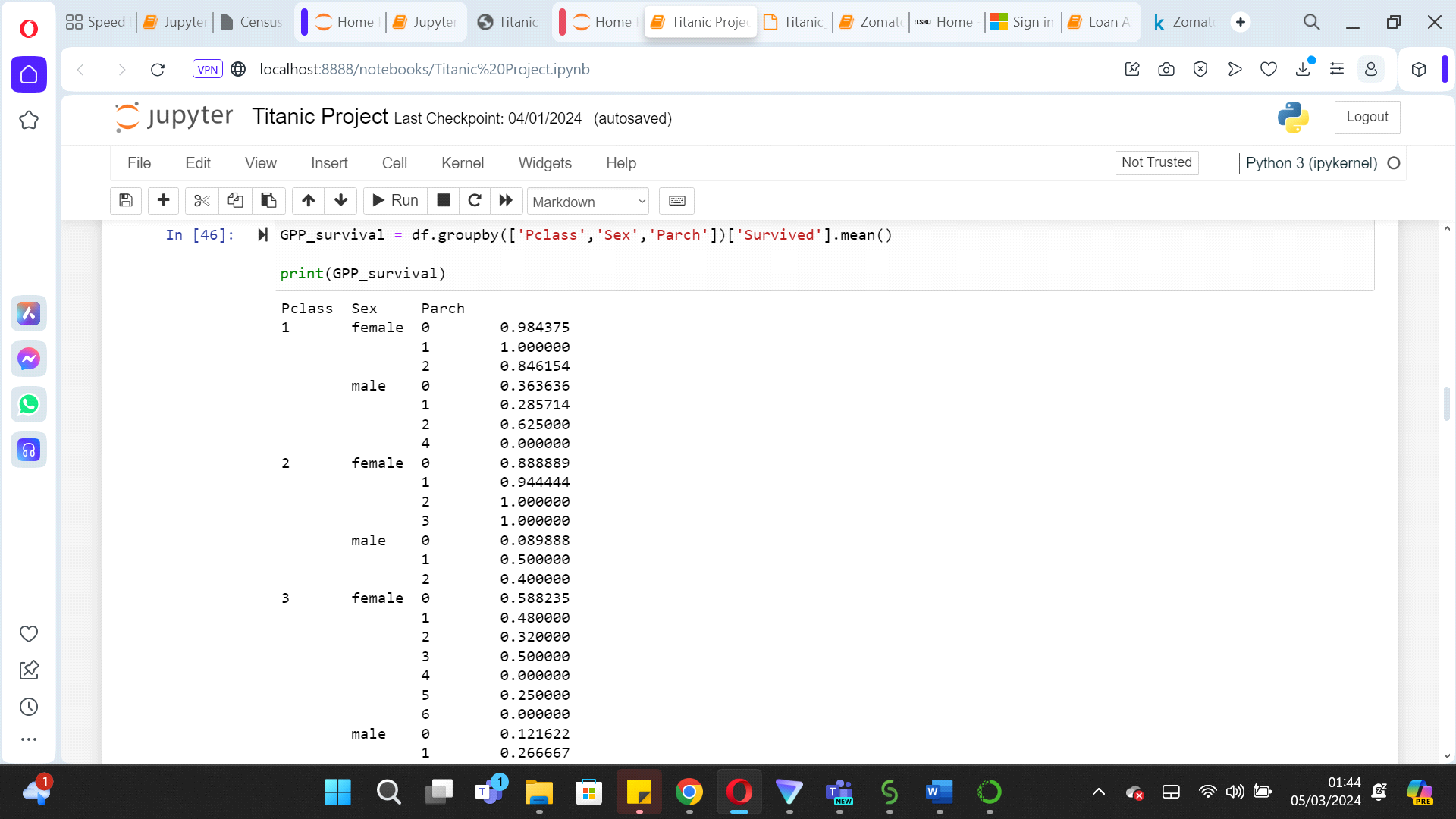


We can derive quite a few insights from this graph:

* The age group 21-30 has the highest number of passengers, with a significant portion not surviving (red).
* The 0-10 age group has a higher survival rate (green) compared to the non-survival rate (red), suggesting that children were more likely to survive.
* As the age increases, particularly after the age of 40, the total number of passengers in each group decreases.
* For the oldest age groups, 71-80 and 81-90, very few passengers are represented, and within these, a very small number survived.

Overall, the graph shows that survival rates were higher for younger passengers, with the rate declining as age increases. Essentially, the graph shows that majority of the passengers did not survive or at least the non-survival rates were higher.

Moreover, looking at the mean of ‘Pclass’, ‘Sex’ & ‘Parch’ on the basis of ‘Survived’



This analysis reveals that first-class women had a significantly high chance of survival, particularly those traveling alone or with one companion.

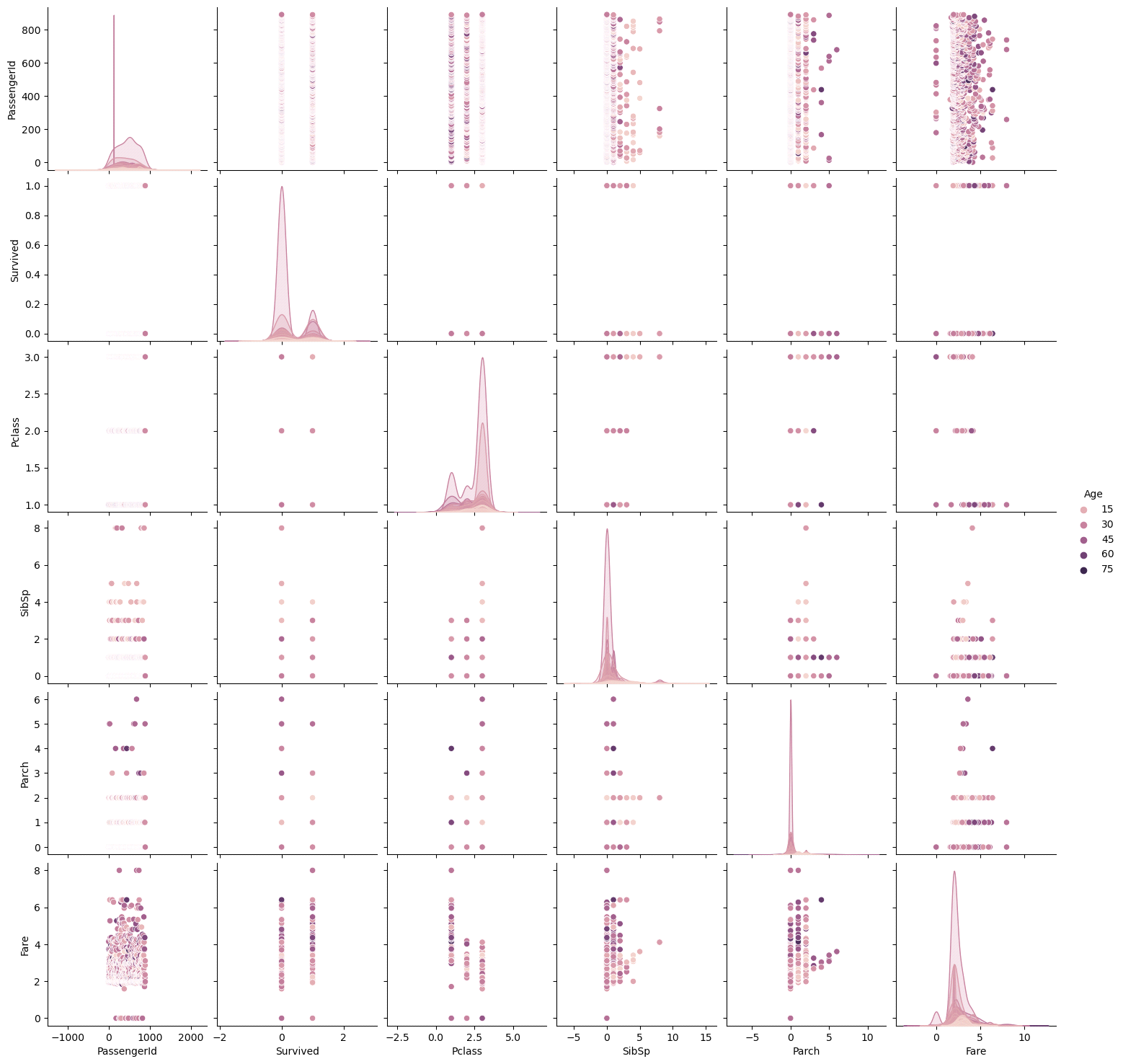
In contrast, men in the same class saw an increase in survival odds when accompanied by at least two others.

In the second class, women with two companions fared better in terms of survival. Men benefitted from having at least one companion, with survival rates of 50% for one companion and 40% for two.

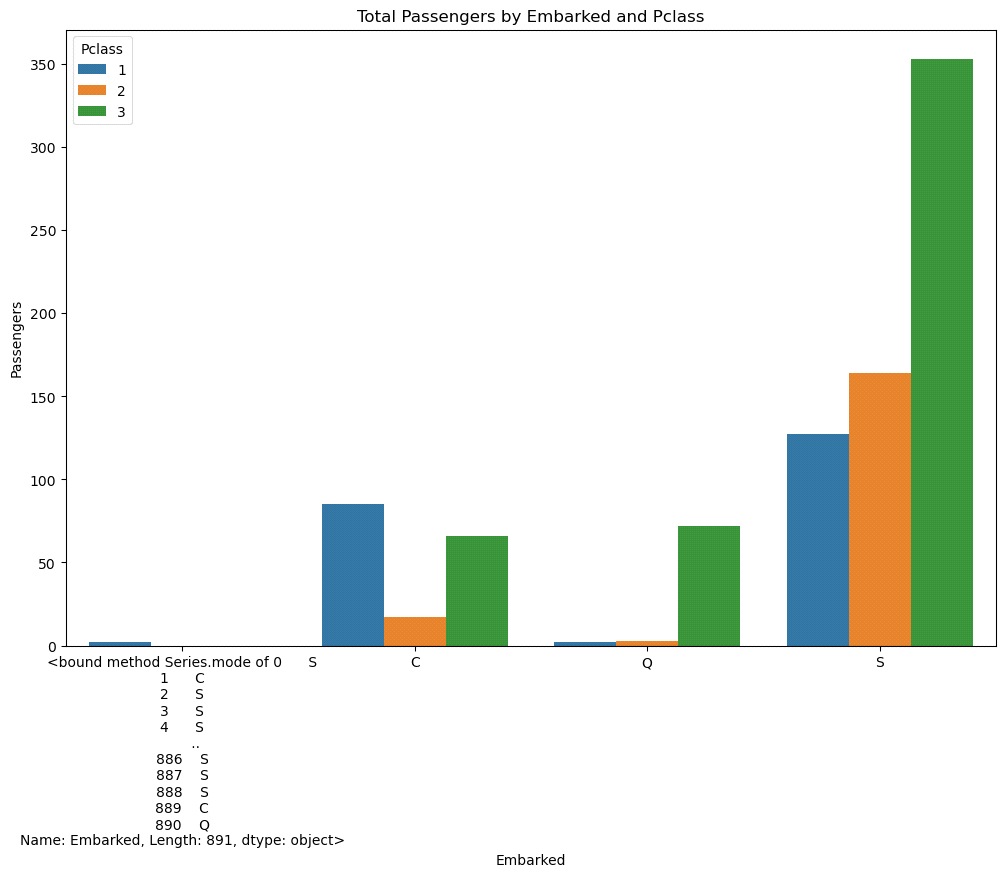
The third class presented a different scenario, where women traveling alone had better survival rates than those with companions. For men in third class, however, having companions was advantageous. This discrepancy could stem from the higher number of women in third class, as indicated by our previous data.

It's important to note the prevalence of single travellers across all classes. The observed higher survival rates for certain groups might be influenced by the smaller numbers of individuals traveling with companions, skewing the statistics.

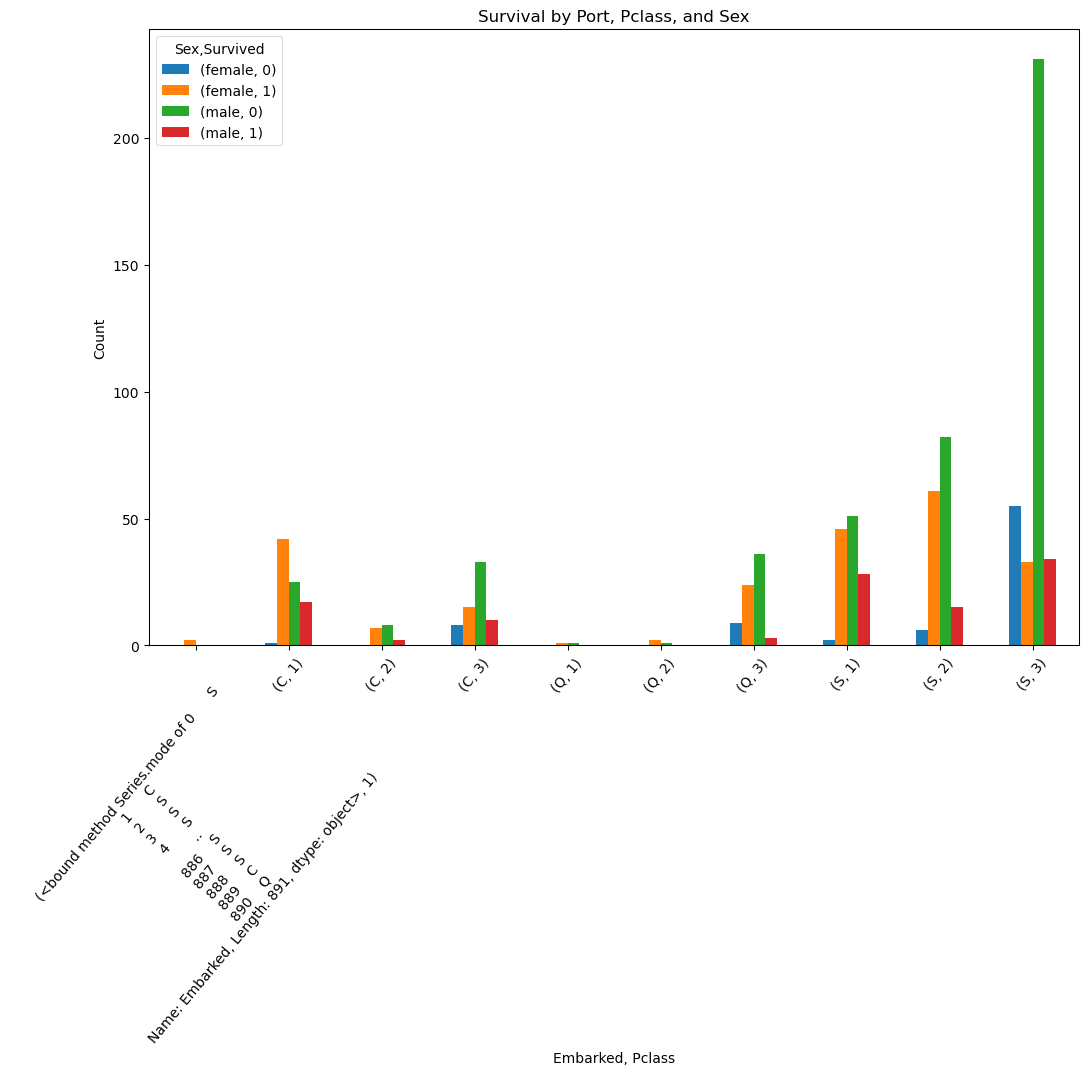
PAIRPLOT



This pair plot was designed to see the relationship between ‘Age’ and the variables.

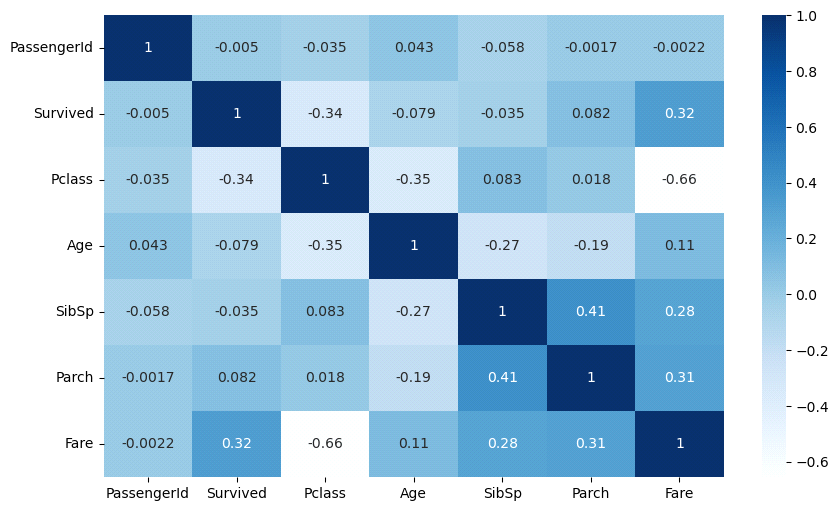


The data indicates that the predominant embarkation point for third-class passengers was Port S. Similarly, first-class passengers mainly boarded at Ports S and C. As for the second class, the majority embarked solely from Port S.

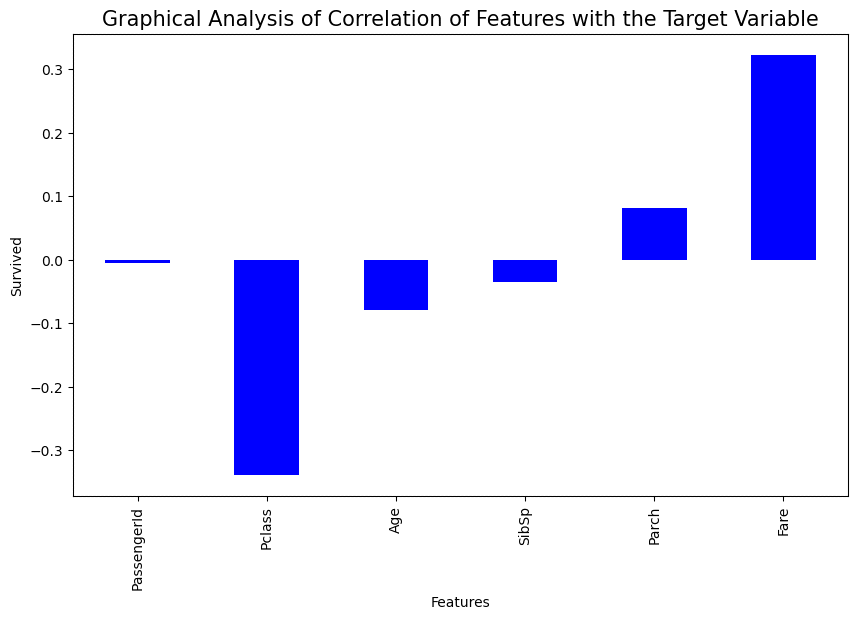
 The chart indicates that women, particularly those embarking from Port S in first and second classes, had better chances of survival, along with Port C's first-class female passengers. Conversely, the graph illustrates those men boarding from Port S in third class experienced substantially lower survival rates.

**Correlation Matrix**

The correlation of coefficient is a relative measure used to understand the strength and direction between two variables. It ranges from -1 to 1 (Negative, Positive or No relation)



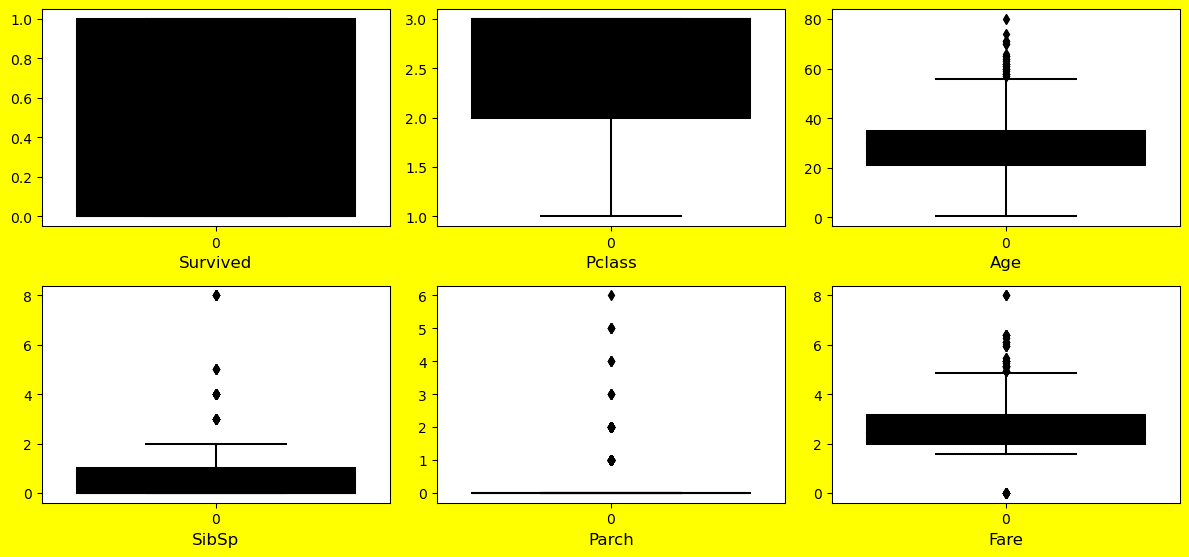
The table above shows some positive relationship between our target variable ‘Survived’ with ‘Fare’ & ‘Pclass’. On the other hand, multicollinearity can be seen in the correlation table between Fare and Parch, Pclass and Fare, SibSp and Fare etc.



The graph above illustrates a stronger correlation between 'Pclass' and 'Fare' and their impact on survival rates.

**Checking for Outliers**

Outliers are extreme values which can be detected using various methods. These extreme values tend to skew the data. The table below shows the presence of outliers for our dataset.

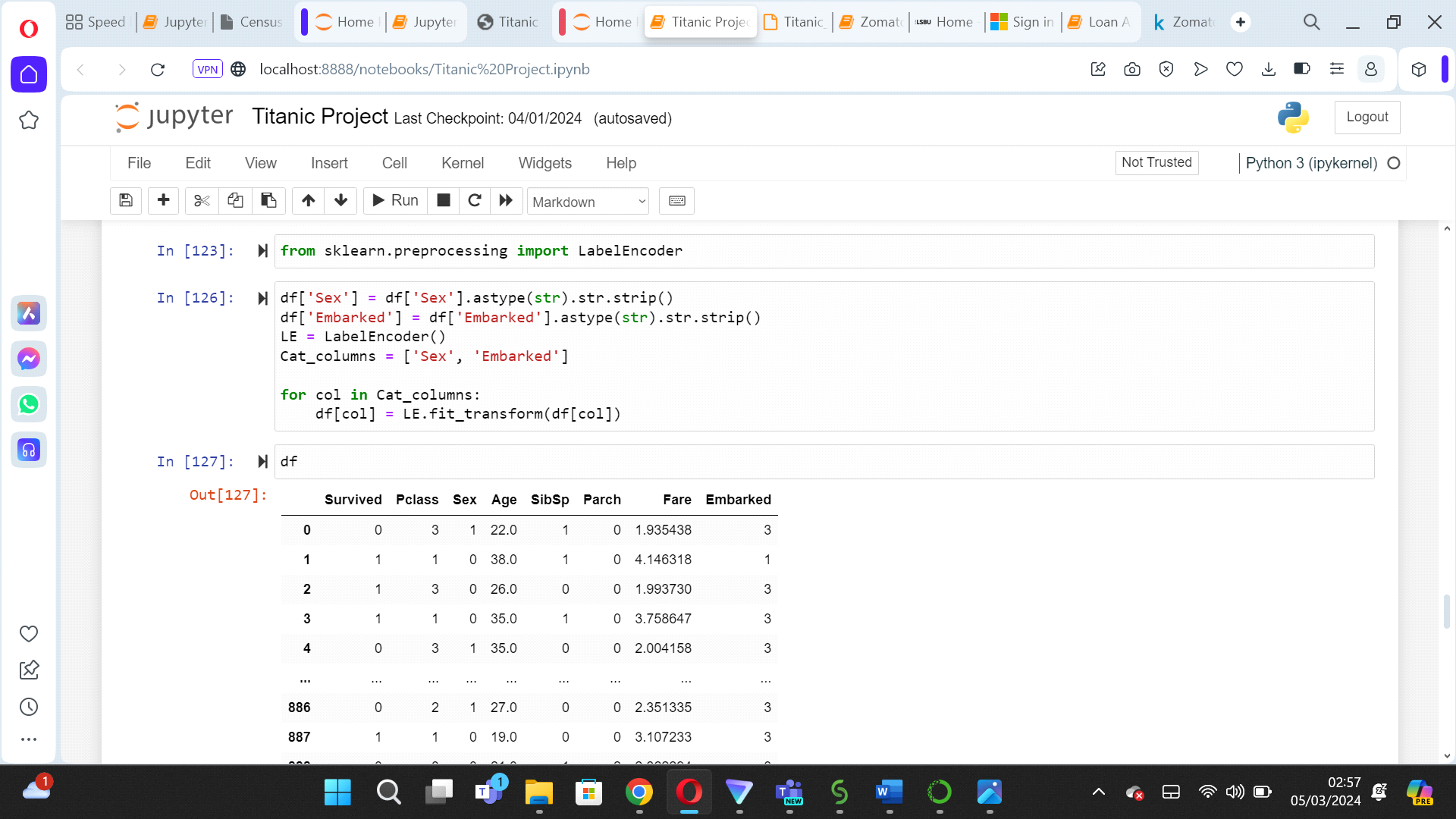


It's observed that the 'Fare' and 'Age' variables exhibit a significant number of outliers. To address this, we can employ z-scores for outlier removal. However, for other variables like 'parch', 'sibsp', and 'pclass', which do hold essential information, the outliers will be retained and because these are categorical variables.

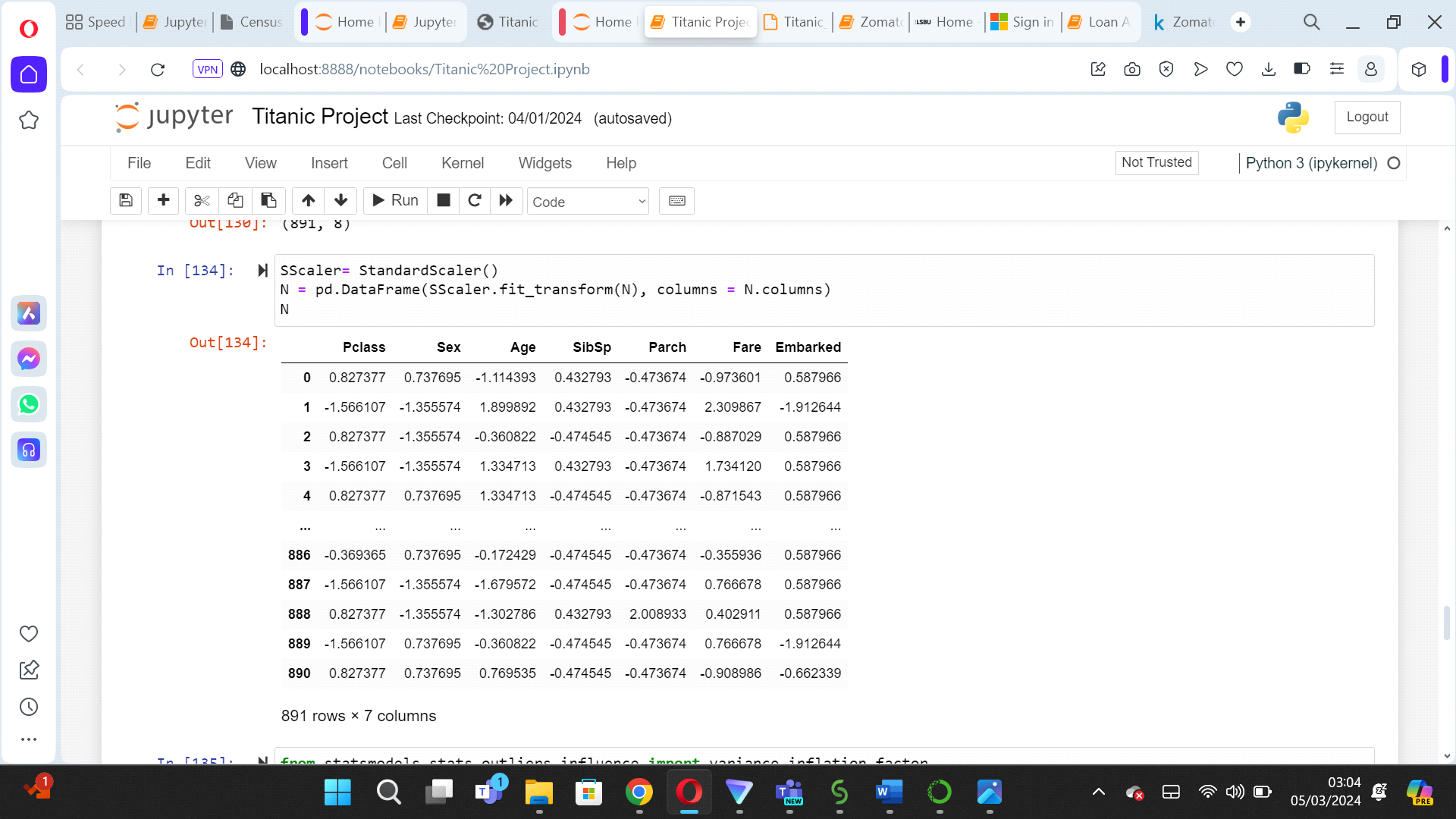
For this dataset particularly, the best approach to address outliers, was to implement the quantile method. This involved calculating the Interquartile Range (IQR) and using it to define upper and lower bounds that help identify outliers. By substituting outliers with the median value for the 'Fare' and 'Age' columns, (which are particularly skewed), we can minimize their effect on the analysis. This method is preferred as it maintains the integrity of the dataset's distribution, especially when the outliers are so pronounced.

**Label Encoder**

We can now apply ‘Label Encoder’ on categorical variables to ensure that we have uniformity across the dataset.

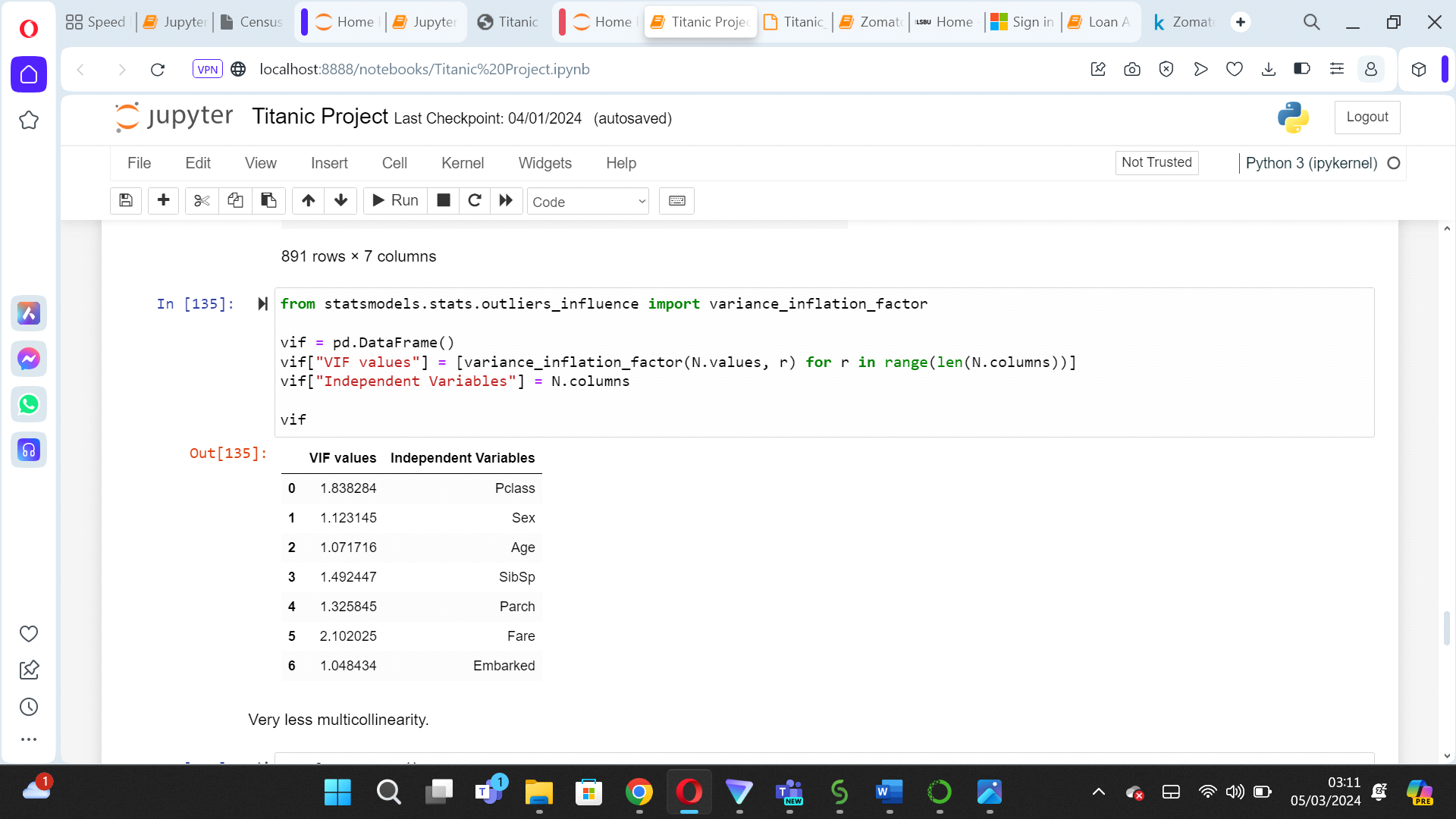


After applying Label Encoder, we can move forward with standard scaling to ensure normalisation as it helps the features to equally contribute to the analyses and having a common scale is beneficial for further Machine Learning parts.

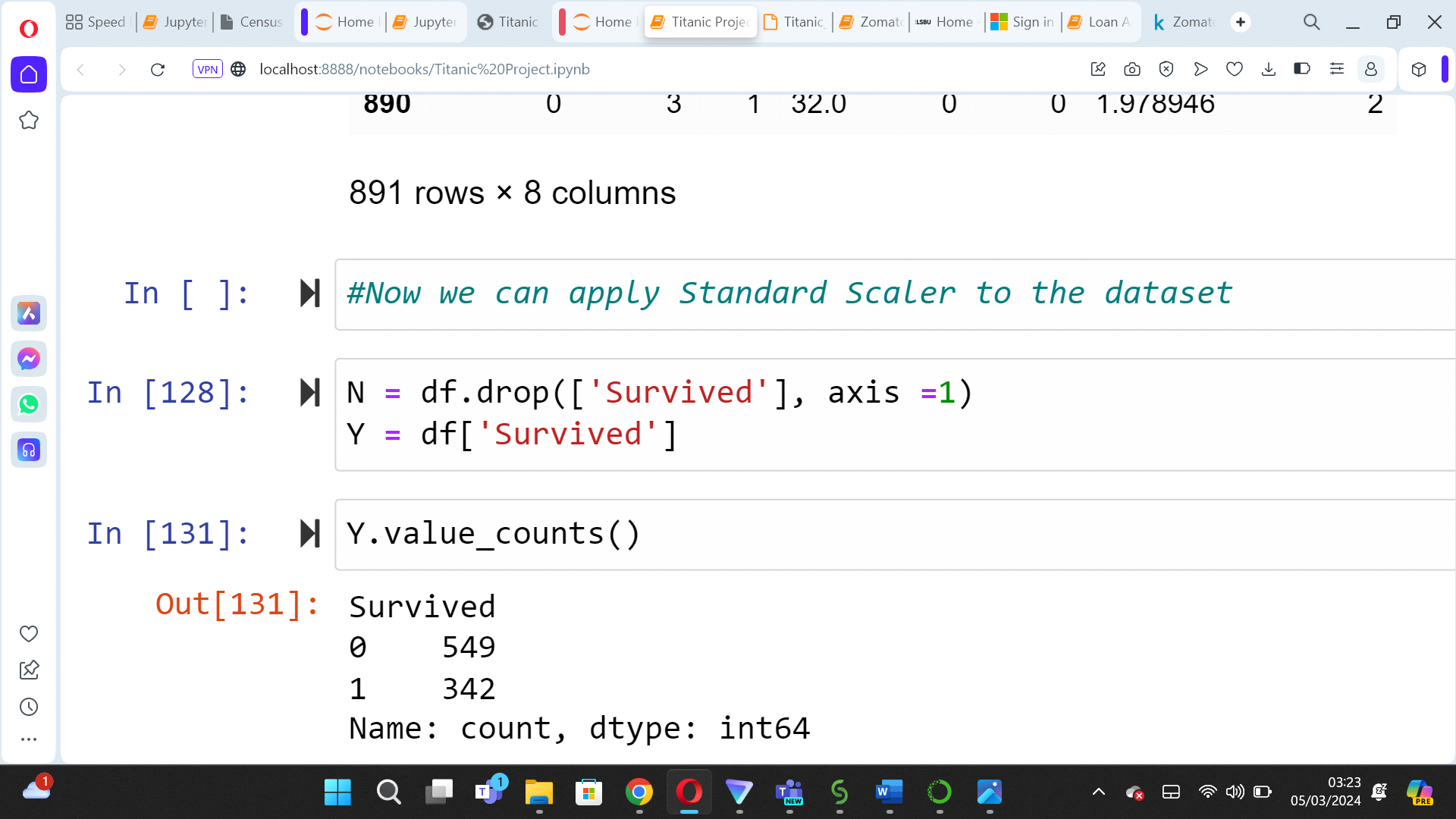


**Testing for Multicollinearity**

While the correlation matrix has already hinted at multicollinearity among variables, we can further validate this through the 'Variance Inflation Factor' (VIF). Fortunately, the VIF scores are coming in under 10, which is within the acceptable range, indicating that our model isn't unduly influenced by multicollinearity and is good to go for further analysis.

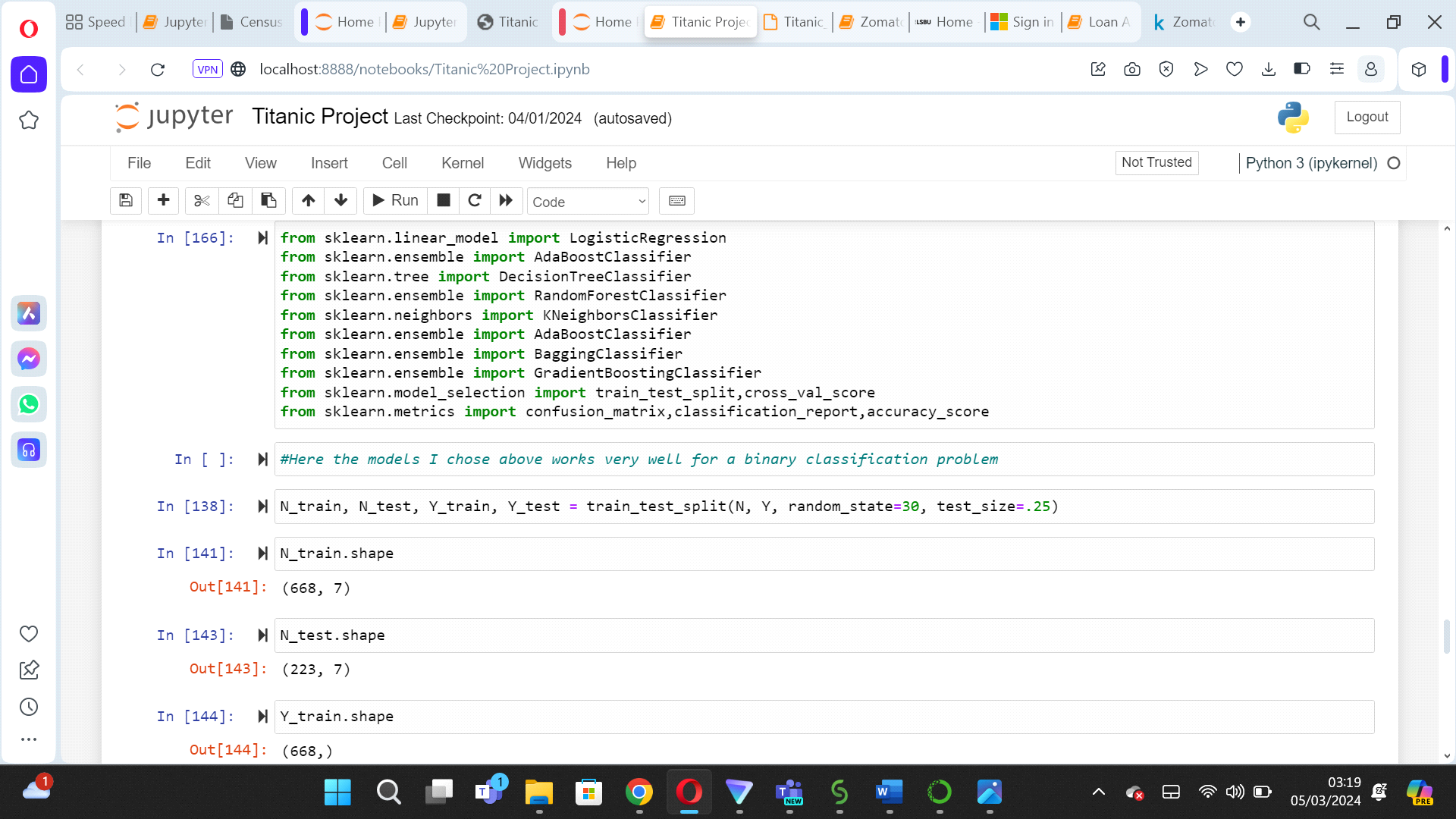


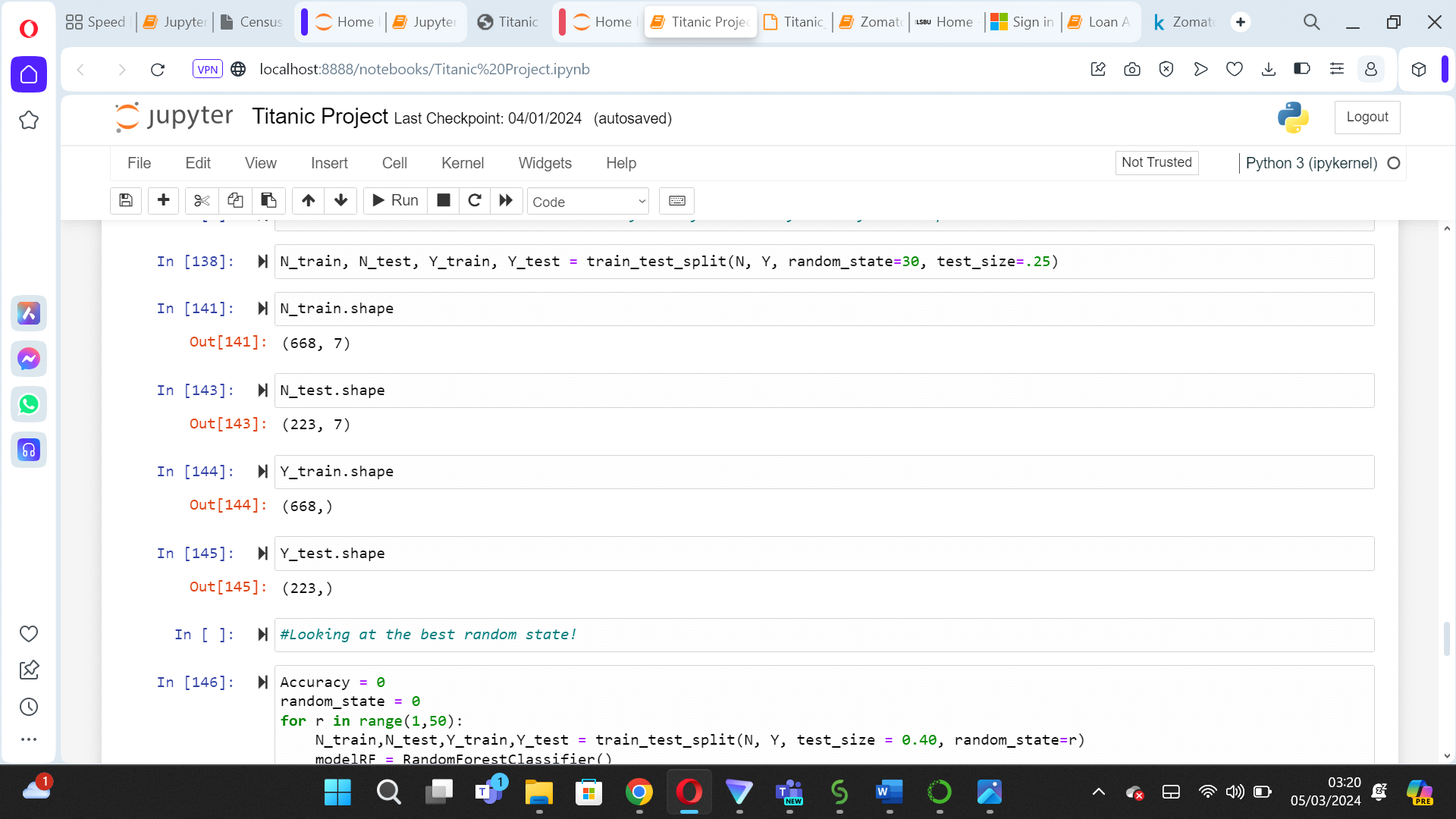
#Beginning the ML Phase



**Machine Learning (Building the Model)**

This segment will concentrate on a supervised learning approach using machine learning models designed for classification. Our objective is to accurately predict whether passengers survived or not. To achieve this, we'll employ the train\_test\_split method, setting aside 25% of the data for testing, while the remaining 75% will be utilized to train our models





Utilising the Machine learning models, the below table is designed to show the results for all the different models including the components required to assess the accuracy for each model.

For 0 (Not Survived)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML Algorithm | Accuracy Score | CV Mean Score | f-1 Score | Recall | Precision |
| GradientBoostingClassifier | 0.8375 | 0.7980 | 0.88 | 0.91 | 0.85 |
| Logistic Regression | 0.8123 | 0.7789 | 0.86 | 0.87 | 0.84 |
| DecisionTreeClassifier | 0.7675 | 0.7609 | 0.82 | 0.81 | 0.82 |
| Bagging Classifier | 0.8235 | 0.7744 | 0.86 | 0.87 | 0.85 |
| Random Forest Classifier | 0.8179 | 0.7800 | 0.86 | 0.85 | 0.86 |
| AdaBoostClassifier | 0.8067 | 07879 | 0.85 | 0.87 | 0.84 |

For 1 (Survived)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML Algorithm | Accuracy Score | CV Mean Score | f-1 Score | Recall | Precision |
| GradientBoostingClassifier | 0.8375 | 0.7980 | 0.76 | 0.71 | 0.81 |
| Logistic Regression | 0.8123 | 0.7789 | 0.73 | 0.71 | 0.76 |
| DecisionTreeClassifier | 0.7675 | 0.7609 | 0.68 | 0.69 | 0.67 |
| Bagging Classifier | 0.8235 | 0.7744 | 0.75 | 0.74 | 0.77 |
| Random Forest Classifier | 0.8179 | 0.7800 | 0.75 | 0.76 | 0.74 |
| AdaBoostClassifier | 0.8067 | 07879 | 0.72 | 0.70 | 0.75 |

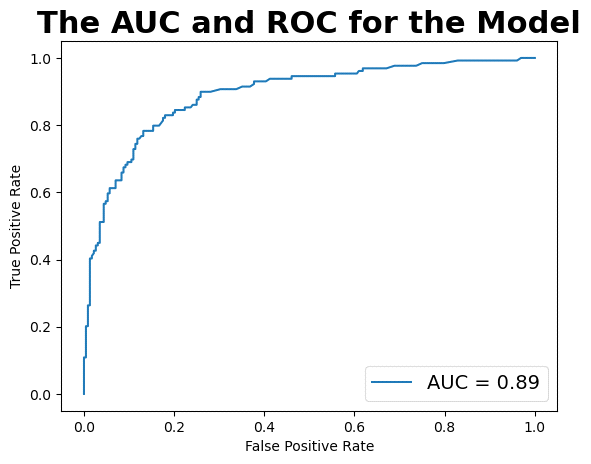
Evaluating all the model’s, Gradient Boosting Classifier has higher scores for F1 score, Recall, Accuracy and Mean Accuracy. Thus, this model will be further used for hyperparameter tunning.

A screenshot of a computer

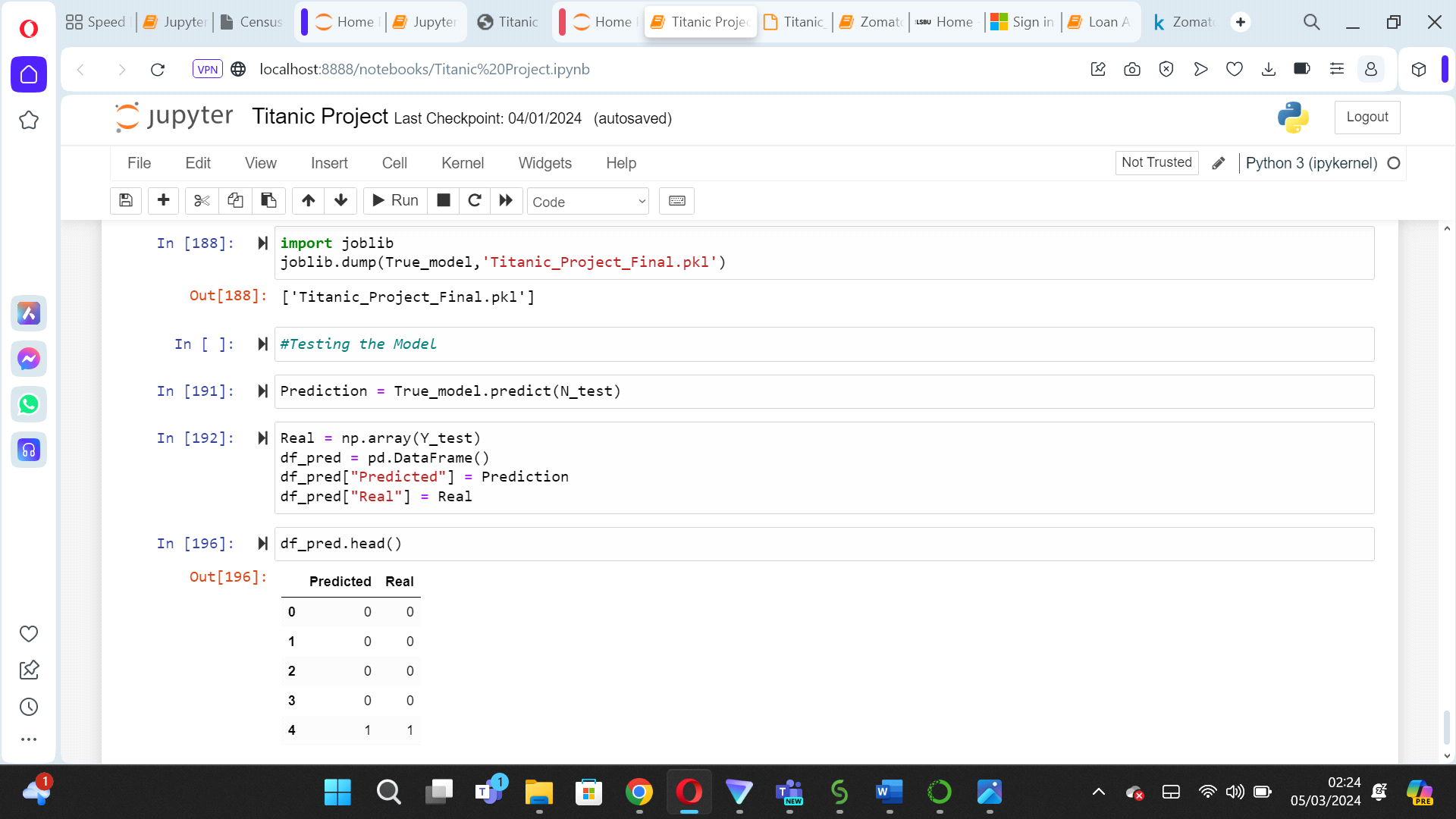
Description automatically generated

A screenshot of a computer

Description automatically generated

After hyperparameter tuning, we observe a modest reduction in the accuracy score, which has dipped to 0.8347 from the pre-tuning score of 0.8375. Additionally, the AUC-ROC score for the Gradient Boosting Classifier stands at 0.89, indicating a robust model performance.

We can now finally use the joblib library to save our final model. Once uploaded, we can reuse the model to make prediction with newer values!



Concluding Remarks:

* Although men outnumbered women on the ship, women had a significantly higher survival rate of approximately 75%, compared to men. Out of 891 passengers, only around (342 passengers) survived the tragedy.
* Survival rates did vary by class: approximately 63% for First Class and 48% for Second Class, showing the significance of socioeconomic status on survival.
* Many passengers boarded from Port S, with most belonging to Third Class. However, those embarking from Port C had higher survival chances, possibly due to the successful rescue of almost all First and Second-Class passengers.
* Women in First and Second Class (Pclass 1 and 2) had good survival rates, regardless of class. Conversely, passengers embarking from Port S in Third Class (Pclass 3) experienced significantly lower survival rates for both genders.