**Anirudh Vunnam**

**06/20/2022**

**Introduction**

In this assignment, the dataset contained information on bank customers of those who accepted a personal loan and those who did not accept the loan. The aim is to build a model that can predict whether customers would take a loan or not. I used a Logistic regression algorithm to build a machine learning model to solve this problem since it can predict a binary class, i.e., Yes or no.

**Data Assessment:**

Immediately after reading data into jupyter notebook using pandas *read\_excel* function, data need to be assessed to check its first few rows, data types, missing data, the shape of the data and have the description of data. This can tell us how clean and tidy the data is. The dataset has 5,000 rows and 14 columns, an output of the *pandas'* *shape* method. The dataset is missing since all columns have "5,000 non-null" values. All columns are in numeric data types, int64 and float64. Numeric data types are readily usable for machine learning algorithms. Besides, I also checked for data duplication; no data duplication was there. Looked into a preview of the data using *describe* function; the function gives an overview of the dataset, for example, mean, median, max, min, std, 25% quantile, and 75% quantile for each numerical column in the dataset. Description of the dataset can tell us a lot about data; listed below is what I drew from the above snippet:

1. Most of the columns had a good data distribution as the data ranges were normal except for the income column, whose 75% percentile is 98 and its max is 224. It is a wide range; also CCAvg column, which is 75%, is at 2.5 and max at ten; this range is wide also. Mortage also has a wide range of 75% = 101, and max = 635. This shows the data has some outliers.
2. The experience column seemed to have negative values as its min value is -3; this raises a data quality issue as an experience of a customer cannot be less than zero.
3. Only zip code seems out of scale range compared to other columns.

**Data Cleaning**

**Converted Experience negative values to positive**

I assumed the negative values could have mistakenly started with the negative indicator and decided to change the values to positive using the python *abs* function as shown below. Apply(), applies the function passed to it along an axis of the DataFrame.

**Explanatory Data Analysis**

In this part, I explore all the dataset variables to discover patterns and check assumptions with the help of summary statistics.

I did divide this section into two parts:

1. Univariate analysis
2. Multivariate analysis.

**Univariate analysis**

Univariate analysis is the analysis of one variable dependently.

For **continuous variables,** I used a histogram to plot them under the distribution of each variable. Histogram discovers underlying distribution of variables (e.g., normal distribution), outliers, and skewness.

Age and Experience columns seemed to have uniform distribution since every value in these columns occurred roughly the same number of times; while Income and CCAvg both had a right-skewed distribution, both had their tail to the right side of the distribution.

For **categorical data,** I used the seaborn counterplot, which shows the distribution of each class in the categorical variable using bars. The family variable occurred to have roughly equal bars, which shows the distribution of each class in the variables was approximately almost equal to each other. The other variables, namely Personal Loan(Target variable), Securities Account, CD Account, and CreditCard, all had a higher proportion of the negative class(0) compared to the positive class(1). The Online variable had a higher proportion of the positive class (1) compared to the other class(0).

**Multivariate Analysis**

Multivariate analysis analyzes more than one variable, mostly a comparison of more than one dependent variable. Using this type of analysis, I compared how variables in the dataset affect each other. I will only discuss significant comparisons.

1. Age and experience had a strong positive correlation; this was shown by the scatter-plot of the two variables, which had a recognizable line with a positive slope.
2. Personal Loan vs. Income showed that higher-income customers are more likely to accept loans compared to those with low incomes.
3. Personal Loan vs. CCAvg showed that customers who spend more on their credit cards are the ones who are likely to accept loans compared to those customers who spend less.
4. Mortgage vs. Personal Loan showed that customers with no or less mortgage are less likely to accept loans than those with a higher mortgage.

**Correlation**

I used a correlation heatmap to check how the variables are correlated to each other. Heatmaps have a color code, which uses different intensities in their color to show how the features are related. Using the Target feature to check for how it is related to other variables, it had a stronger positive correlation with Income, CCAvg, and CD Account variables than other variables. It also had a weak negative correlation with experience and Age variables.

From the heatmap, Age and Experience shows that they are very strongly correlated against each other, as shown by the heatmap color code.

**Modeling**

The problem was classification supervised machine learning; thus, suitable classification algorithms can be used to model this data, for example, Logistic regression, Random regressor, naive Bayes, KNN, and Random Forest Classifiers I used Logistic regression for this problem.

Logistic Regression

Is an algorithm used to solve classification-oriented problems? It can predict binary classes, multiple classes, or even ordinal classes (ordered) [6]. The problem in question is binary, and this model is fit for the problem. Linear Regression and Logistic Regression are quite similar. Linear Regression is used for regression problems, while Logistic Regression is utilized for classification problems. Instead of fitting a regression line, we fit a Sigmoid Function ("S" shaped logistic function) in logistic regression, which functions as an activation function in machine learning to add non-linearity to a linear function and may map any real value between 0 and 1 of the Sigmoid Function.

Model performance

I dropped some non-useful variables before modeling the data; zip Code and Id columns don't carry much weight in modeling.

I used accuracy, confusion-matrix, and classification report metrics to check the model's performance.

1. Accuracy → is the simplest metric; it is the number of correct predictions divided by the total number of forecasts multiplied by 100 percent. The model had a training accuracy of 0.948 and a testing accuracy of 0.944; the two accuracies are roughly equal. The model is neither over-fitting nor under-fitting. Since Our Target variable was imbalanced, we cannot use accuracy alone to judge our model. Confusion matrix and classification in such a case always come in handy.
2. Confusion-matrix → is a 2d-table of (actual and predicted) that describes the performance of a classification model. Using this table, we can tell how many values are predicted correctly and which are also predicted wrongly. The confusion matrix has four terms:

True Positive (TP) → is a metric that shows how many positive class samples the model accurately predicted.

True Negative (TN) → is a metric that shows how many negative class samples the model accurately predicted.

False Positive (FP) → shows the number of how many negative class samples the model predicted wrongly.

False Negative (FN) → shows the number of how many positive class samples the model predicted wrongly.

This model had 1113 True Positives, 68 True Negative, 17 False Positives, and 52 False Negatives. From this metric, the model predicts the high-proportion more than the less-proportion class.

1. Classification Report: this is another classification metric that shows the model's precision, recall, and F1 score.

Precision → is the number of positive class predictions compared to the number of true and false positives. The model had a precision score of 0.80 for those who would take a personal loan.

Recall → is the number of true positive class predictions compared to the total number of True Positive and False-negative.

F1 score → is the weighted harmonic mean of precision and recall. The model's anticipated performance will be better if the F1 score is near 1.0. The f1 score for the model for the positive class is 0.66, while the F1 score for the negative class is 0.97, the model performance is roughly good, and it is also imbalanced since the target class was imbalanced; the model works perfectly with the high proportion class.[

Important features for the model

Three most important features:

1. Education
2. CD Account
3. Family

Age was the variable that had the most negative influence on the model.

Conclusion

I built a bank personal Loan prediction model, which predicts whether a customer would take a loan or not. Since the problem was a classification, the Logistic Regression algorithm was a good choice for building the model. The model performance was not perfect, but it had a good performance.

Recommendation

The model performance was not good enough, and some improvement is needed to improve the model performance. The dataset needs to have a more balance target class to better the model performance. This imbalance was a big issue; as I tried to remove outliers from the dataset, I lost more data on the less-proportion target class. The use of logistic regression model is utilized in determining the status of loan approval and therefore whether the customer would take a loan or not and the method will essentially assist the loan credibility prediction systems in making better decisions in terms of whether to reject or approve loans from customers however this method is only limited to prediction and hence incorporation with other techniques outperforming the performance of this model needs to be tested and eventually implemented for the domain to assist the banks further.

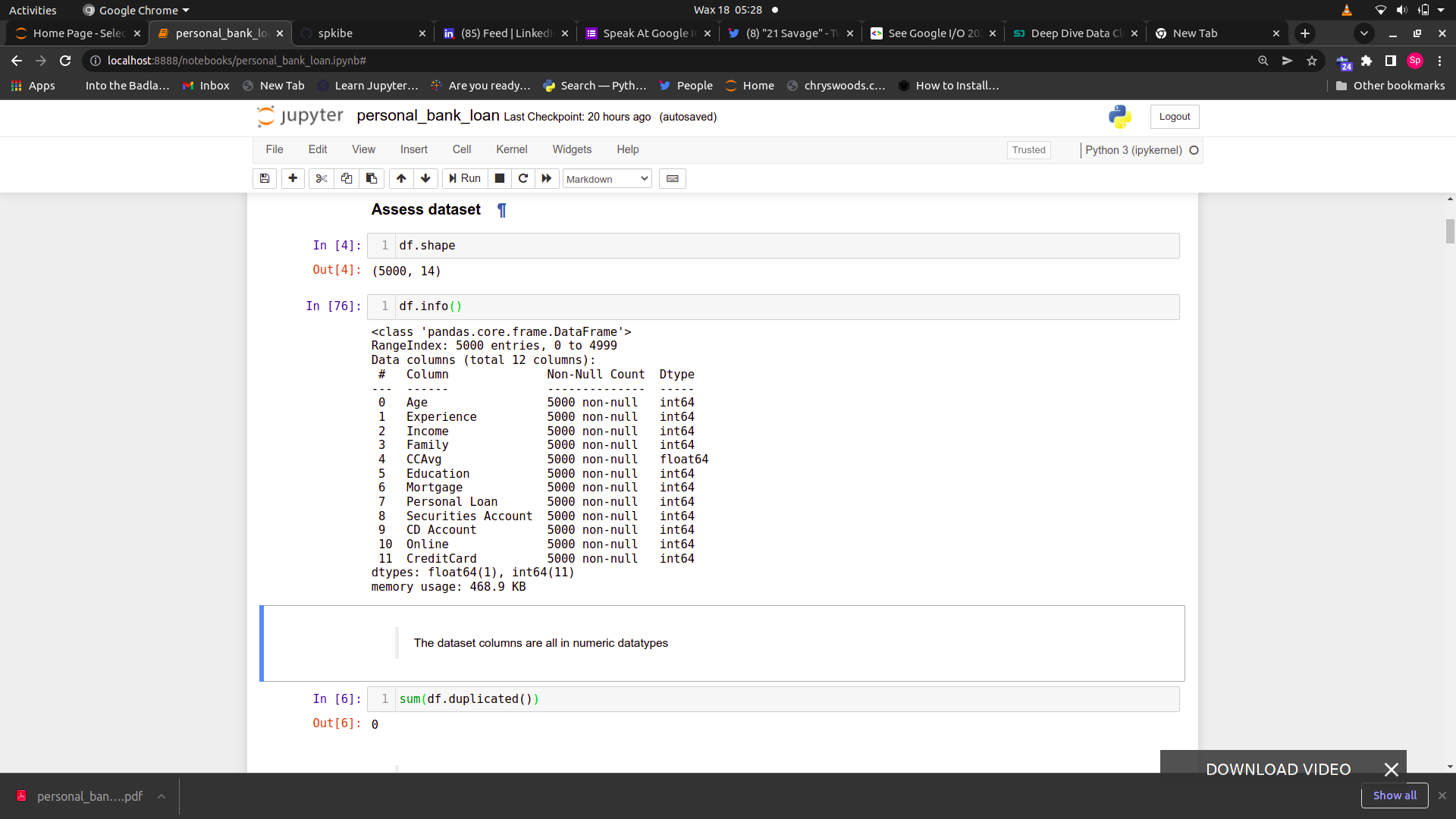
**Reference**

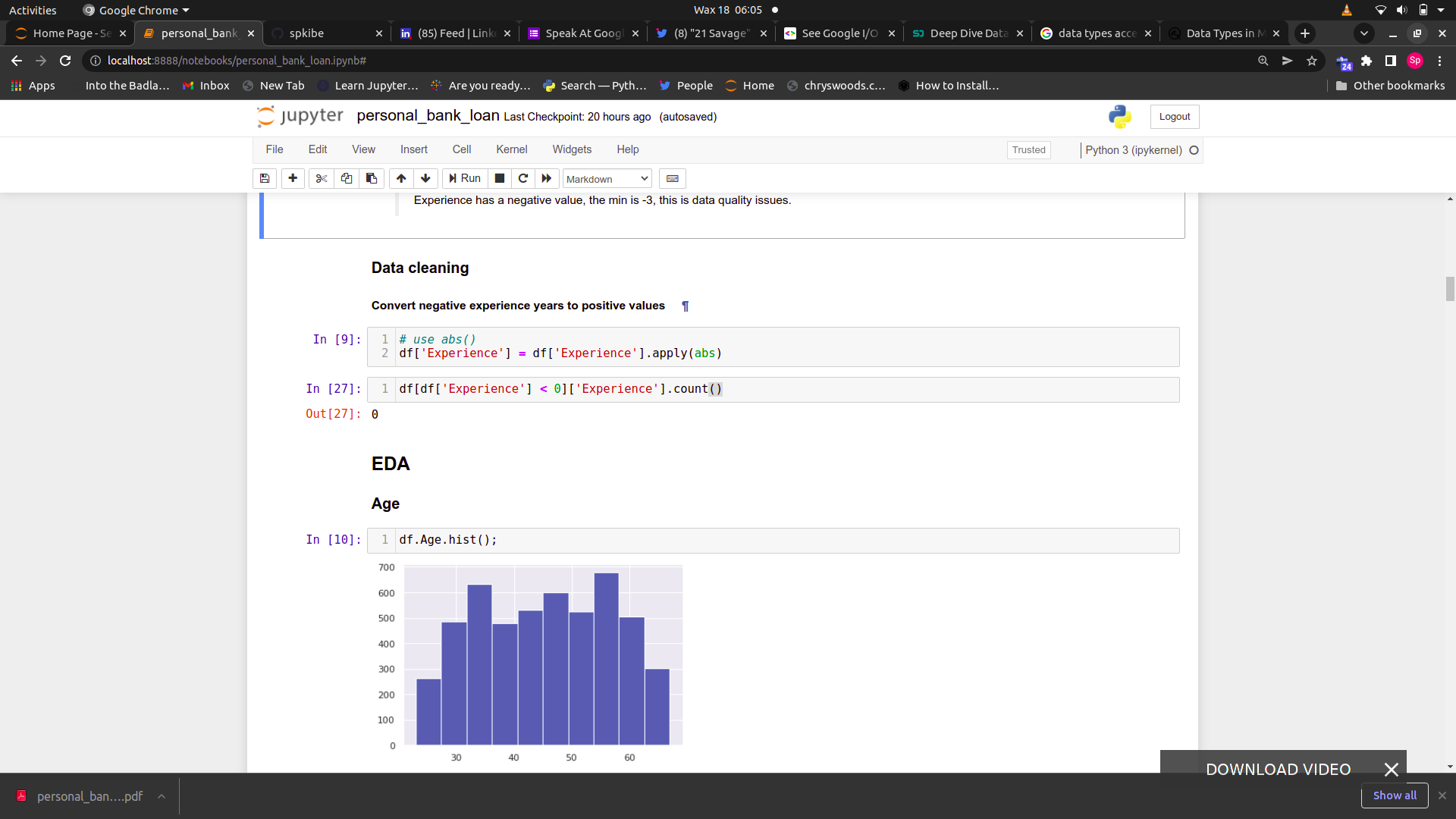
Divate, Y., Rana, P., & Chavan, P. (2021). Loan Approval Prediction Using Machine Learning. <https://5y1.org/download/8f5f1c064fd94364f9746b092aa8e68e.pdf>

Itoo, F., & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), 1503-1511. <https://link.springer.com/article/10.1007/s41870-020-00430-y>

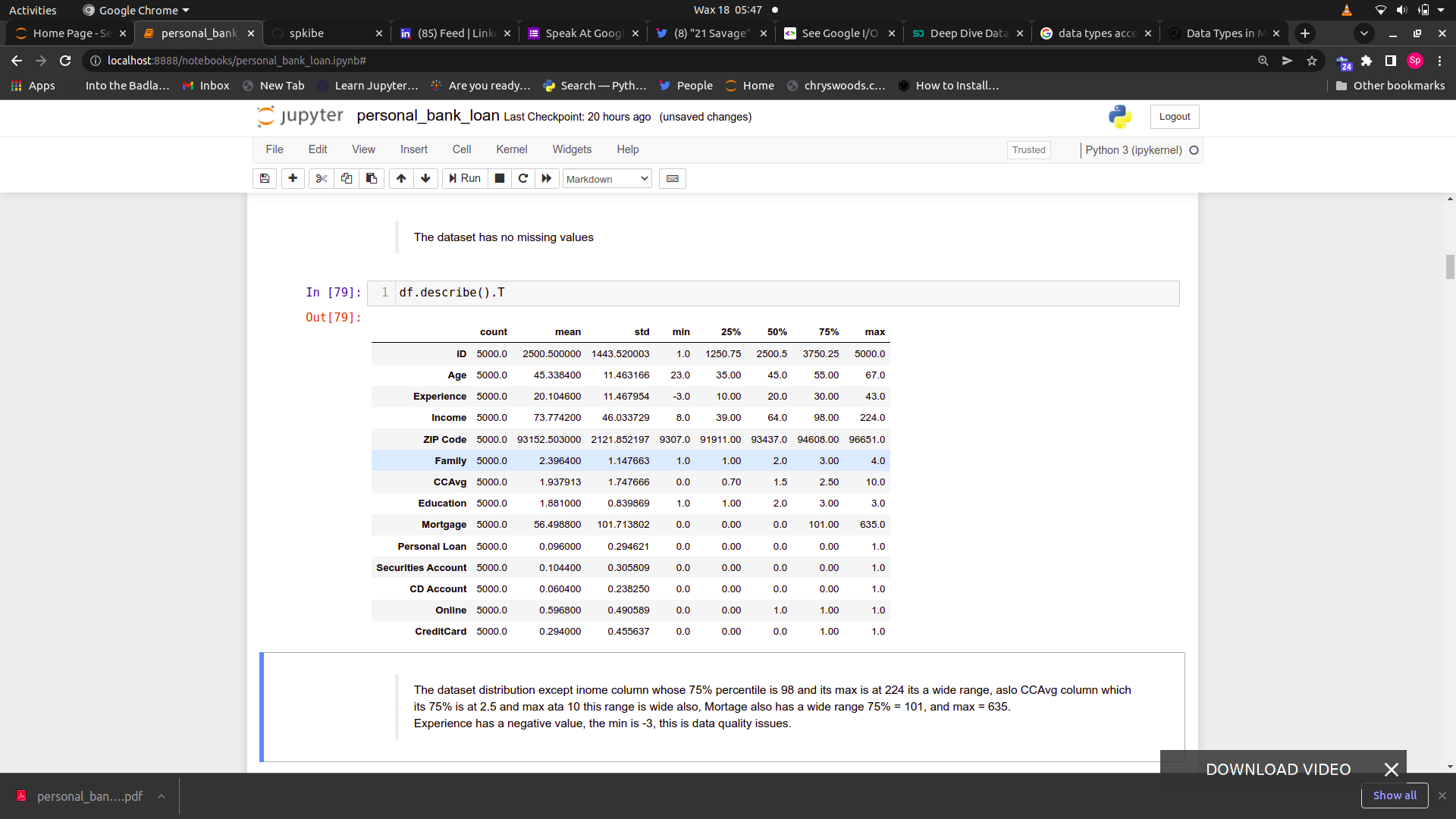
TVS, J. (2021). Predicting the Loan Status using Logistic Regression and Binary Tree <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3769854>

**APPENDIXES**

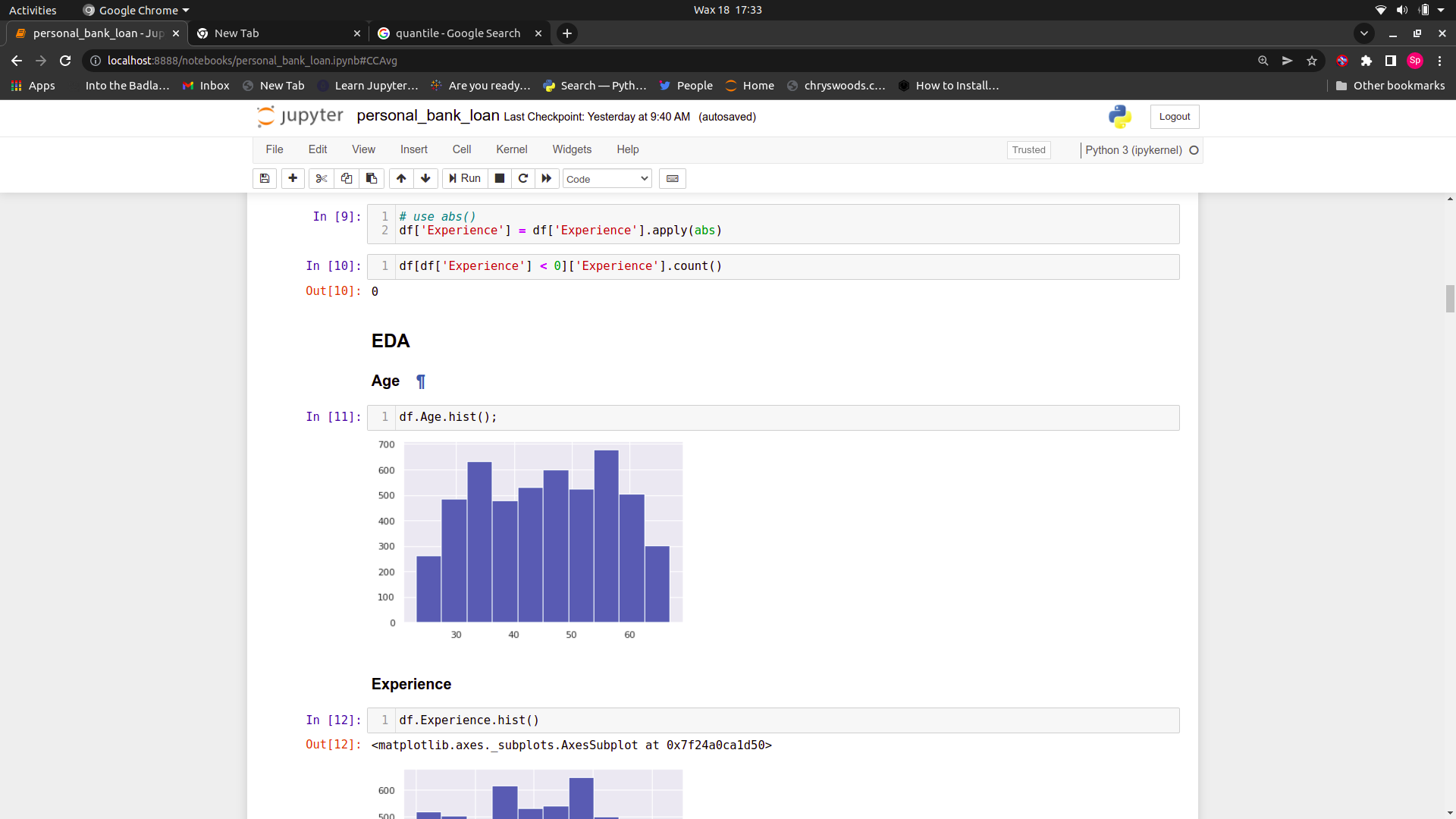
****

****

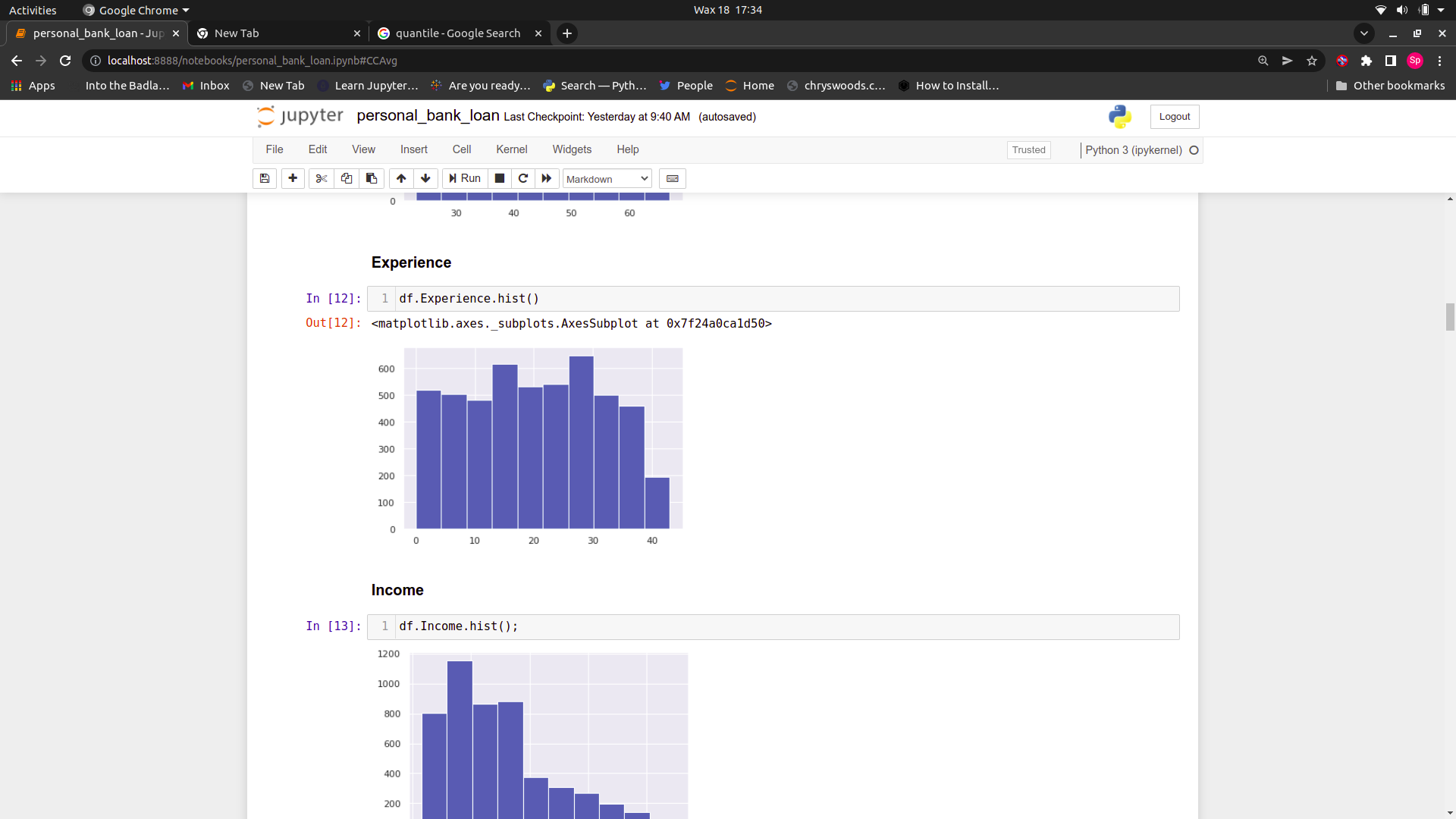
**Data Description**

****

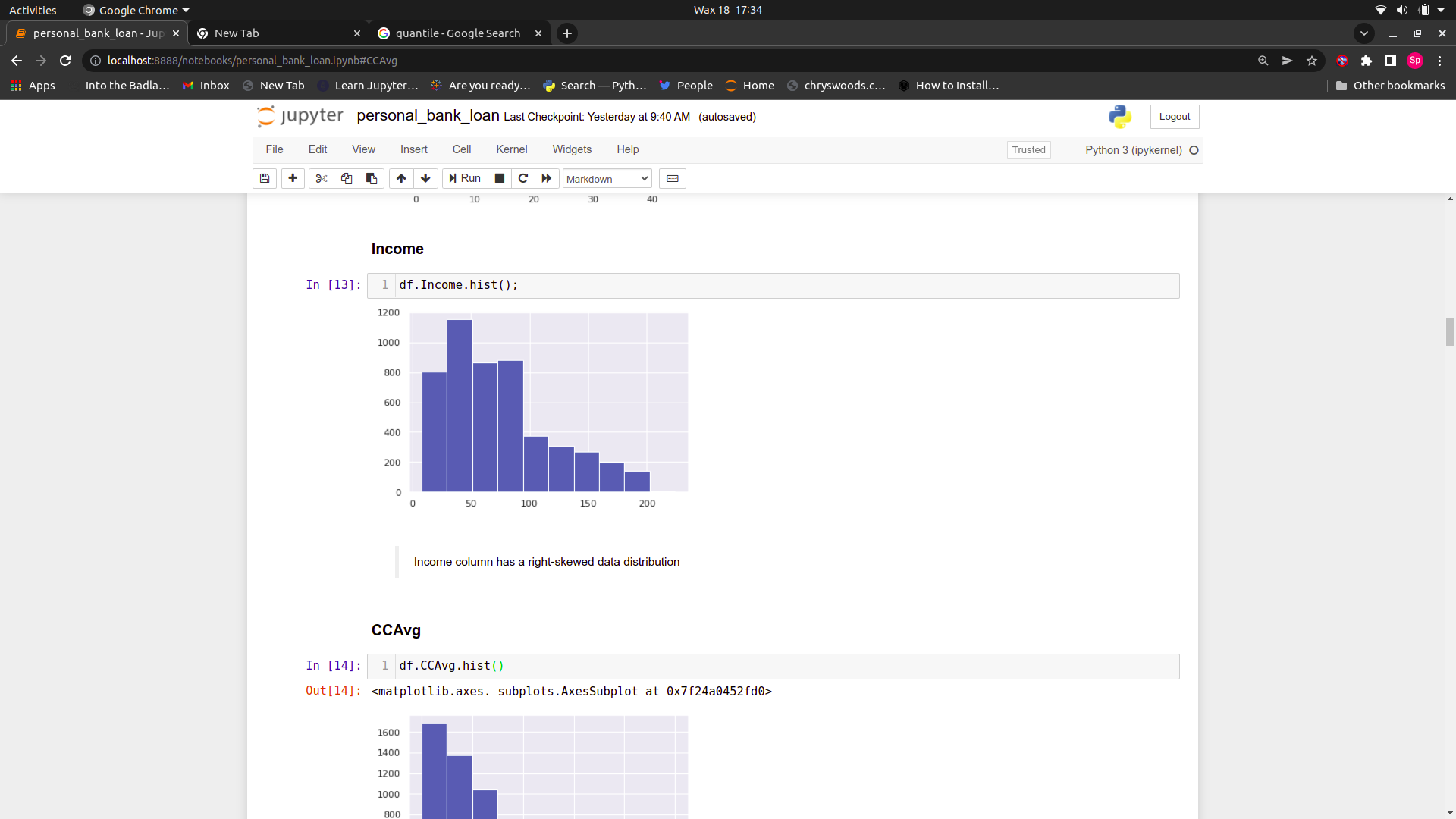
**Histogram of Age**

****

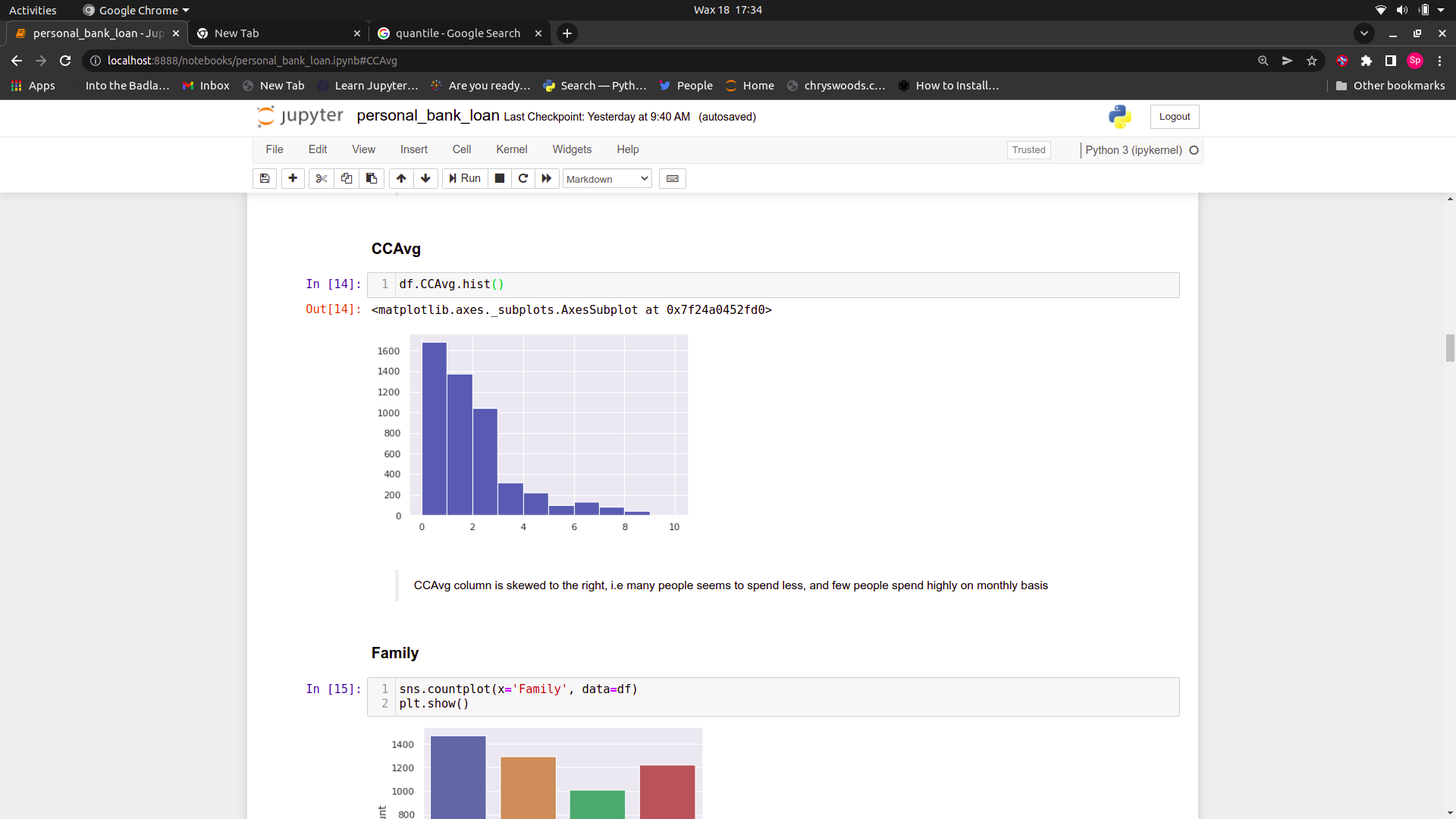
**Histogram of Experience**

****

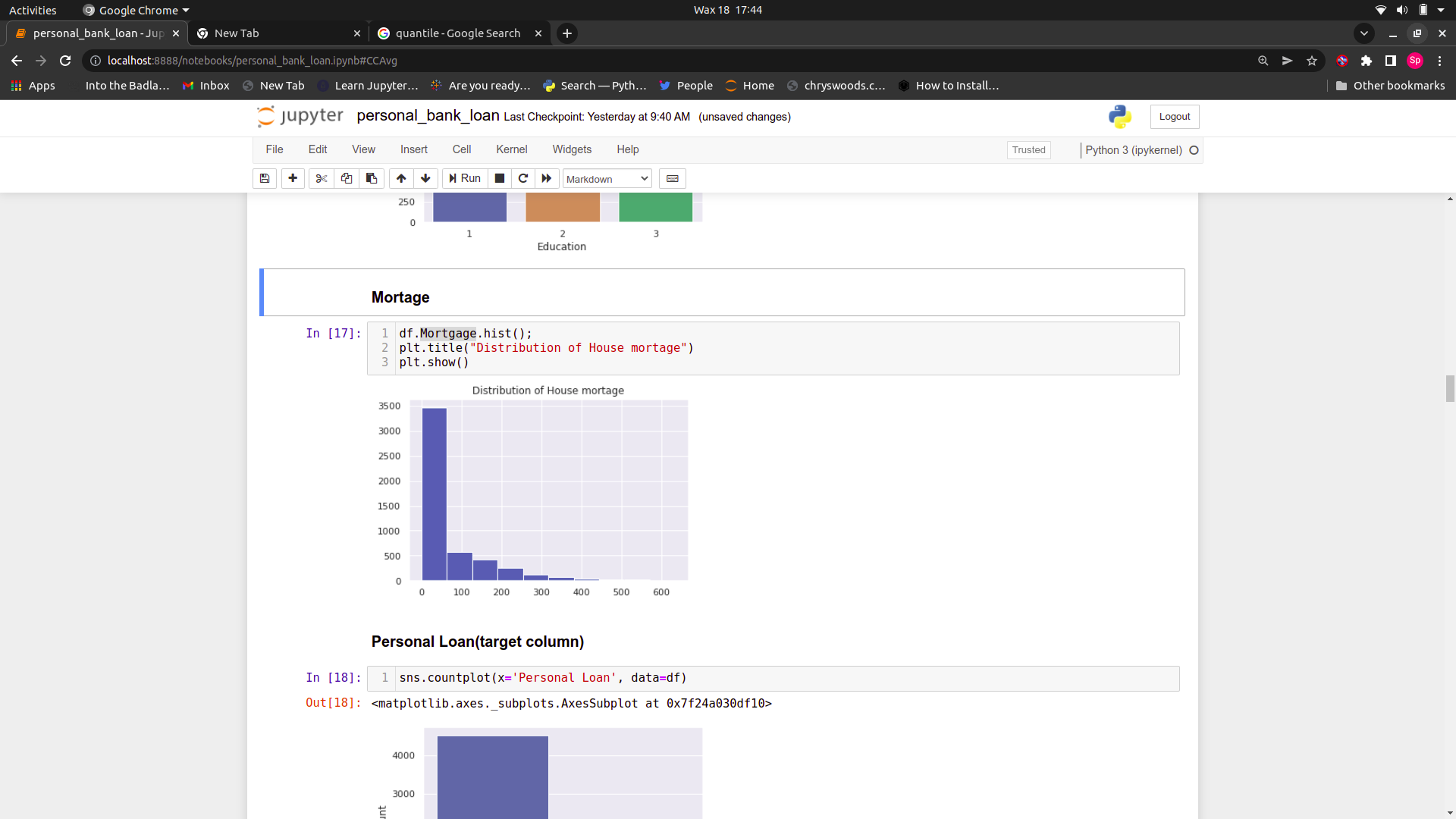
**Histogram of Income**

****

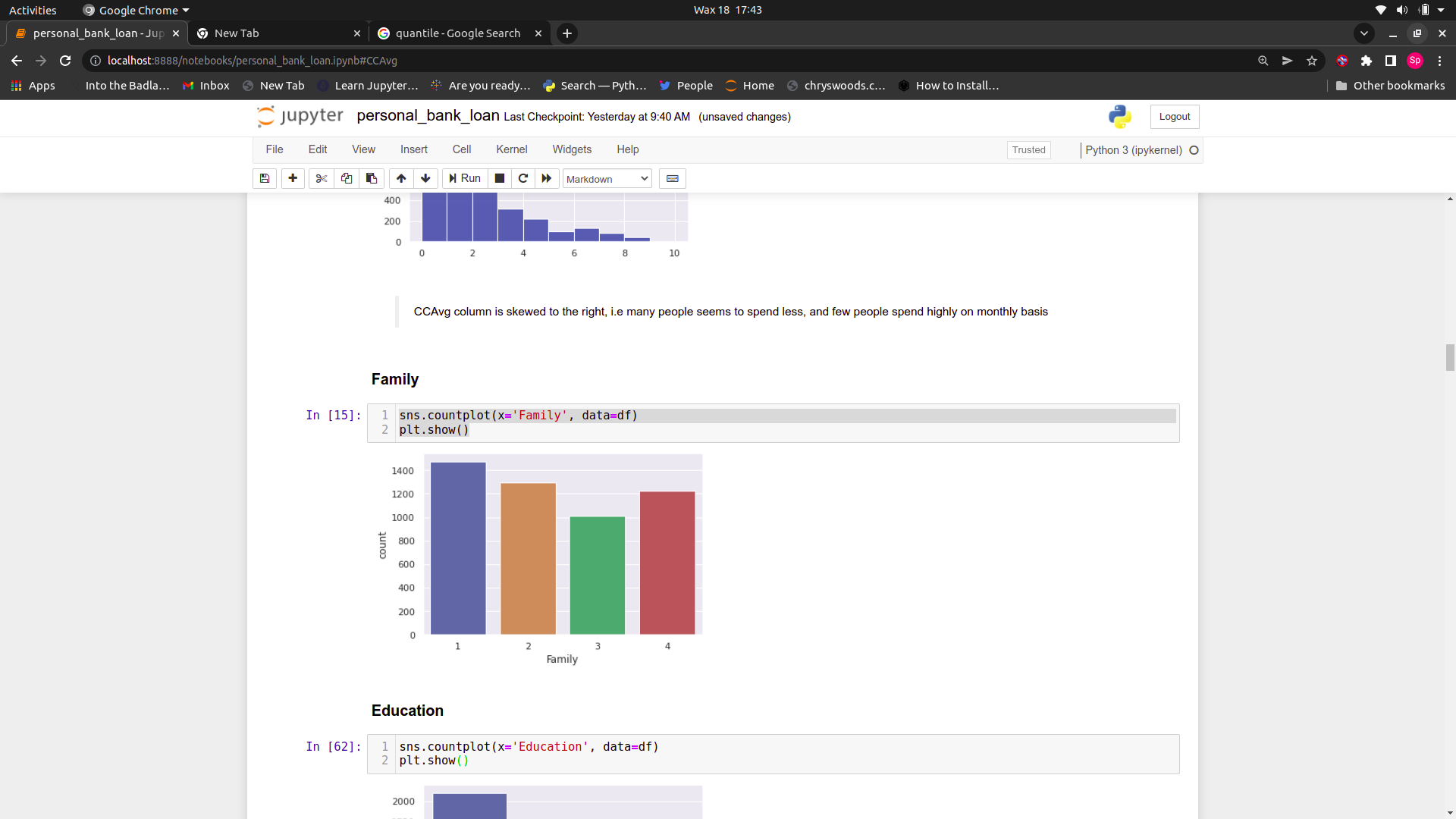
**Histogram of CCAvg**

****

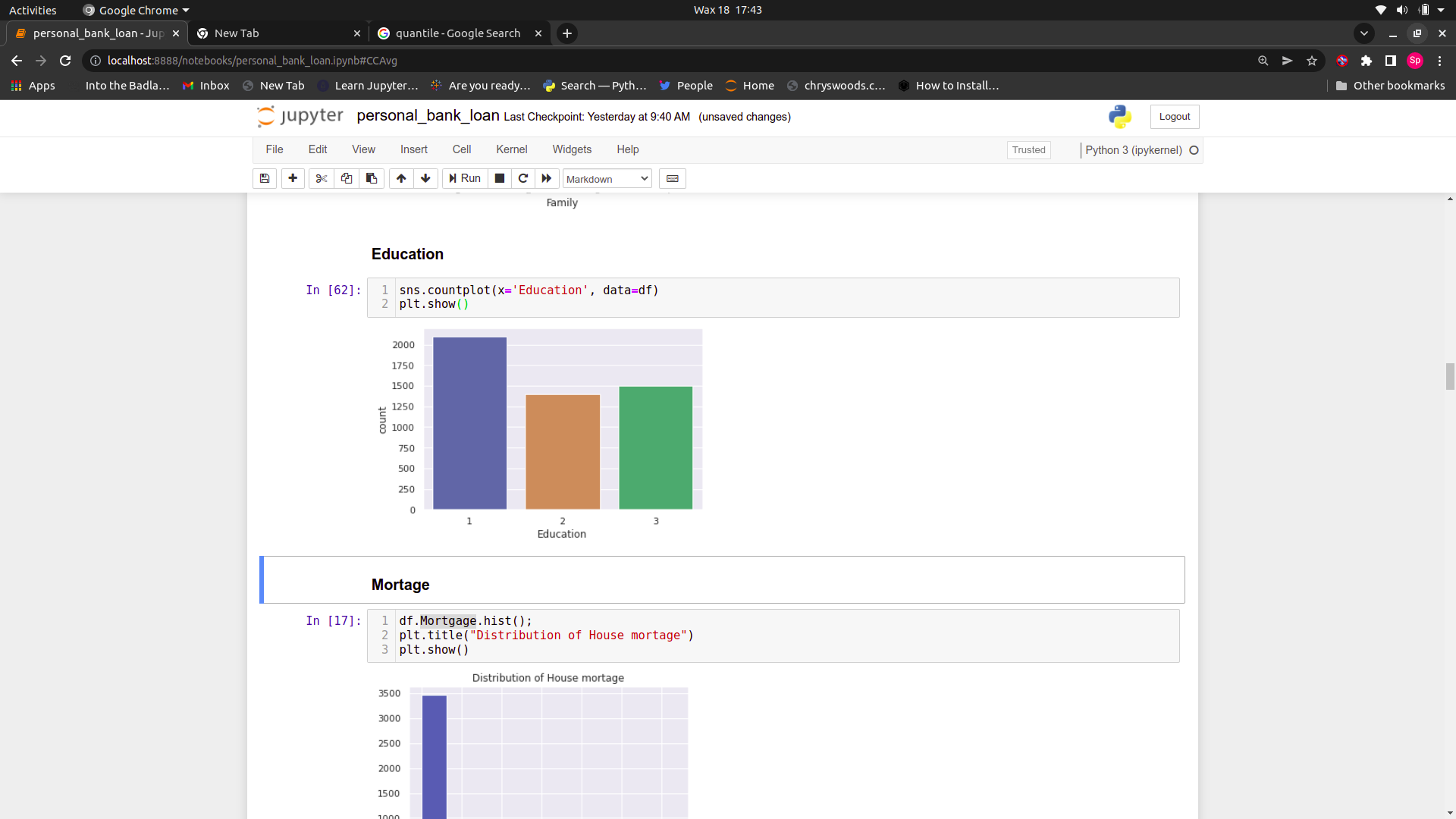
**Histogram of Mortgage**

****

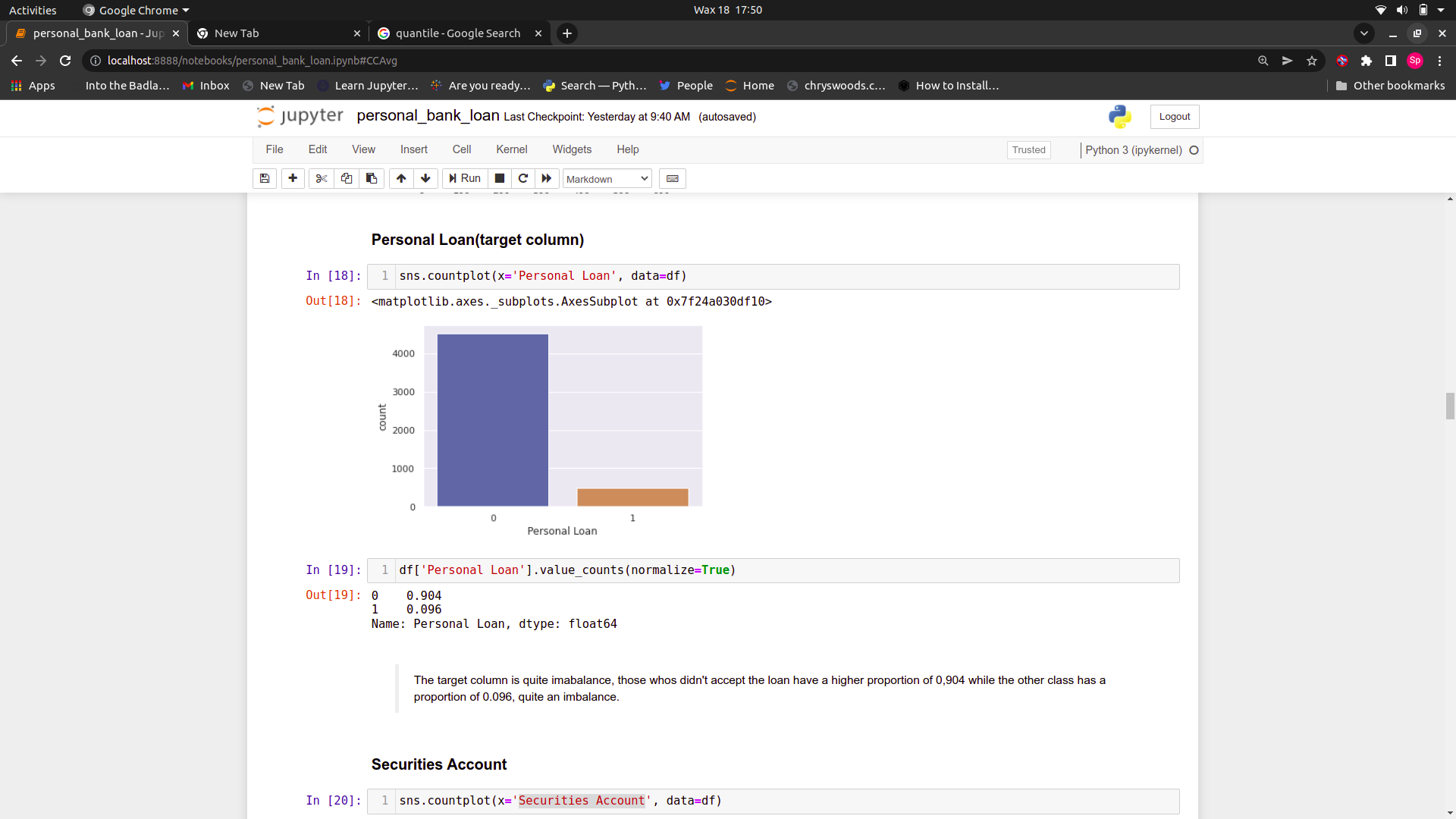
**Countplot of Family**

****

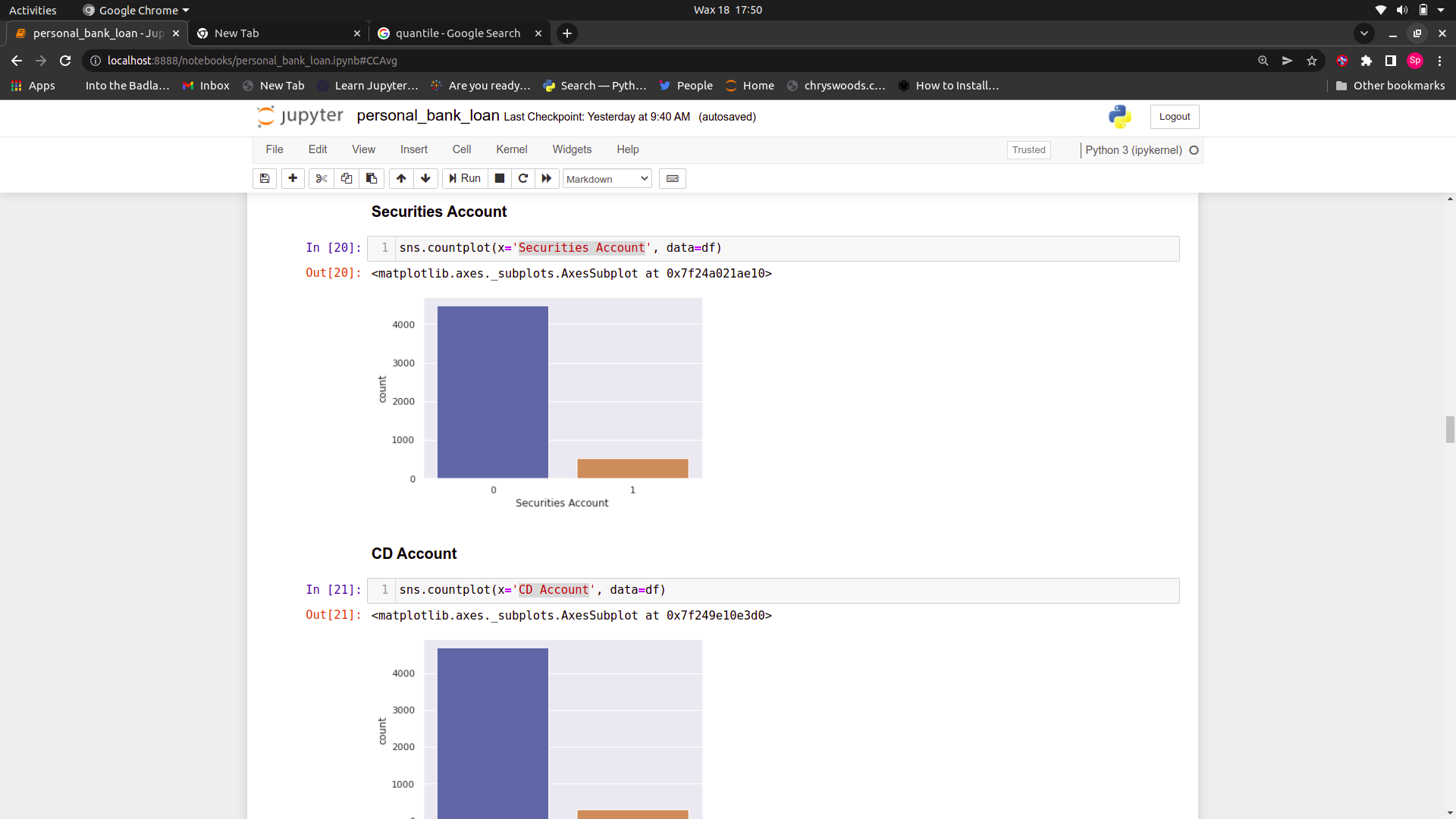
**Countplot of Education**

****

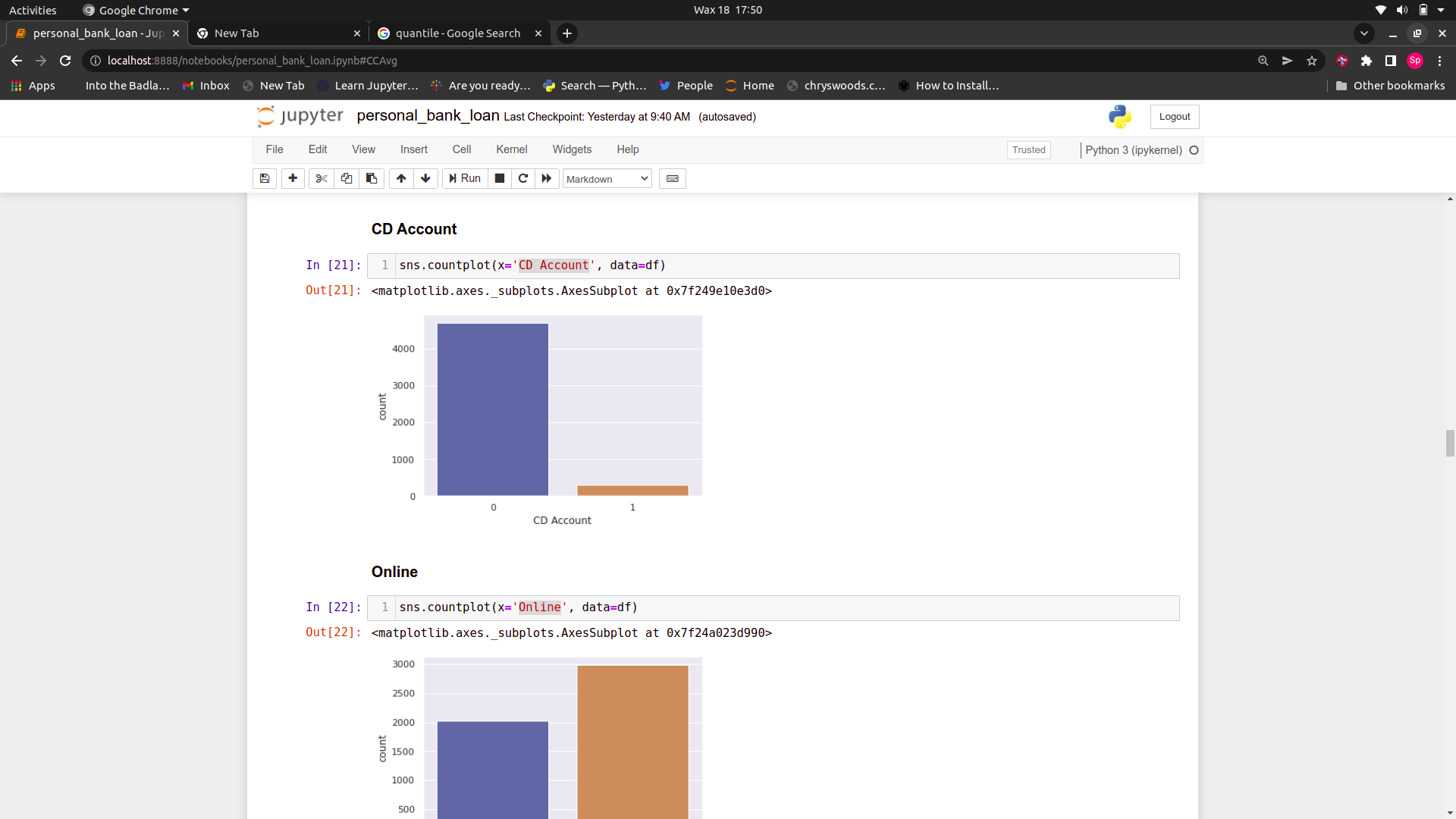
**Countplot of Personal Loan(Target variable)**

****

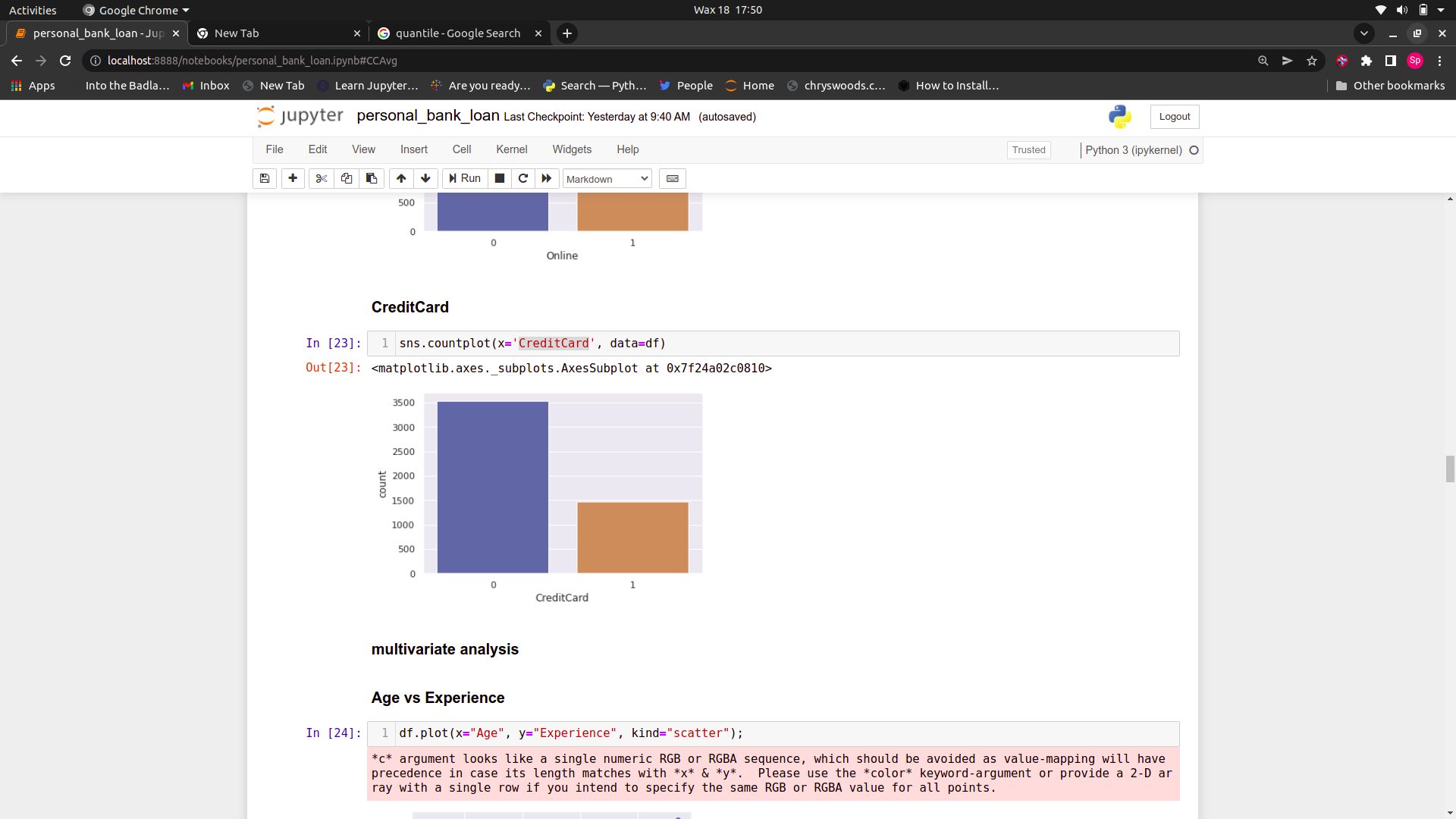
**Countplot of Securities Account.**

****

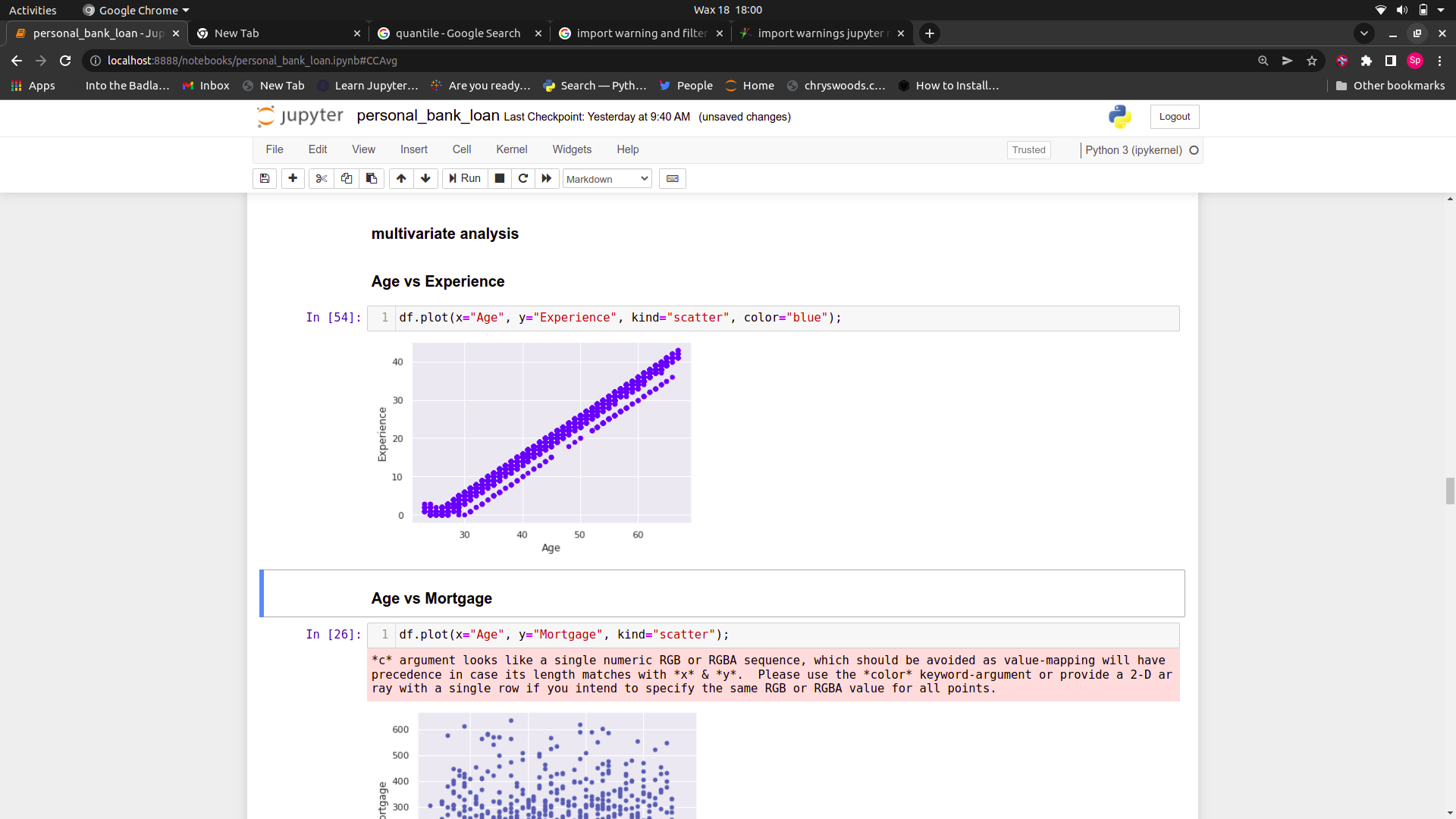
**Countplot of CD Account**

****

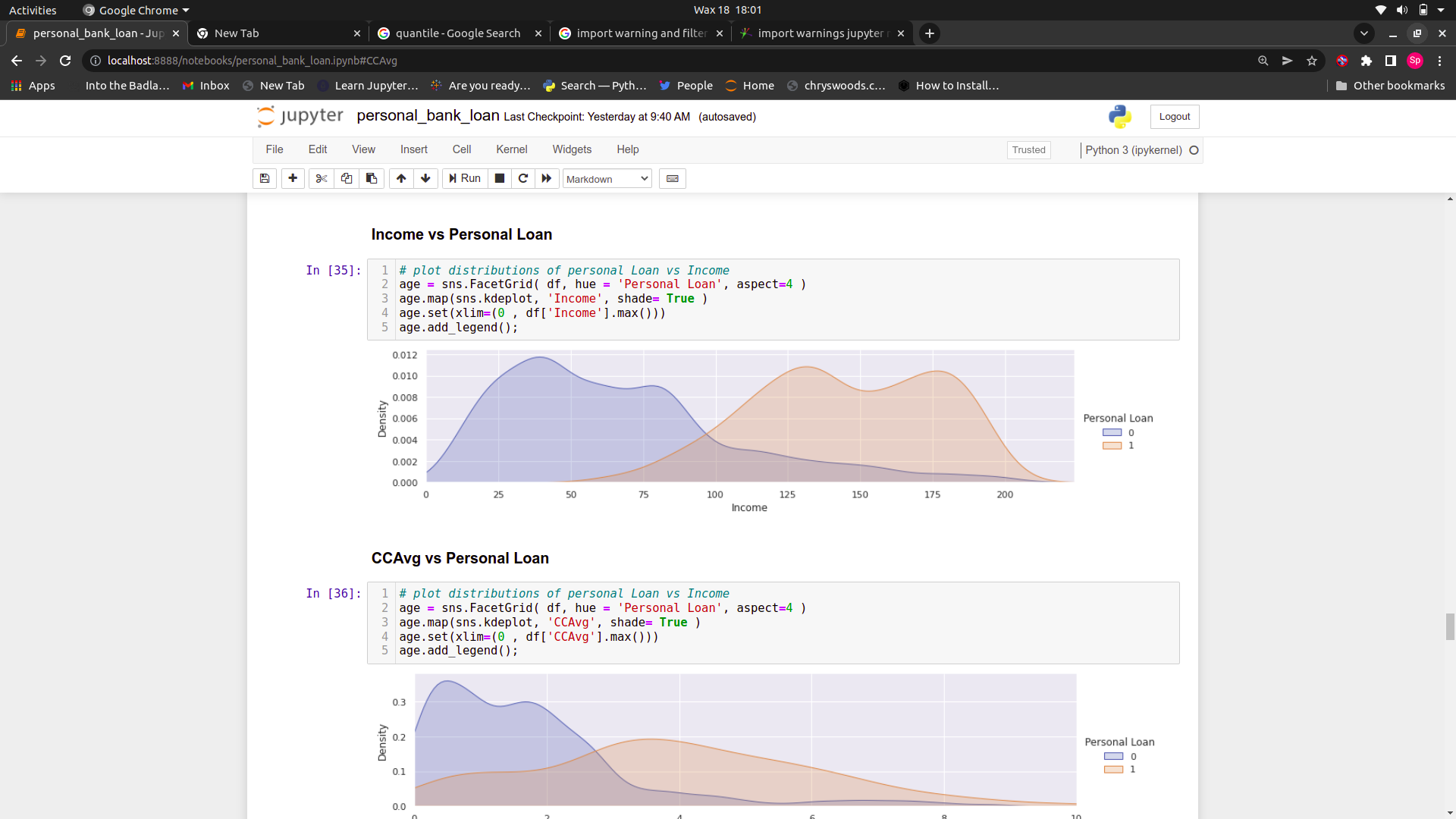
**Countplot of Creditcard.**

****

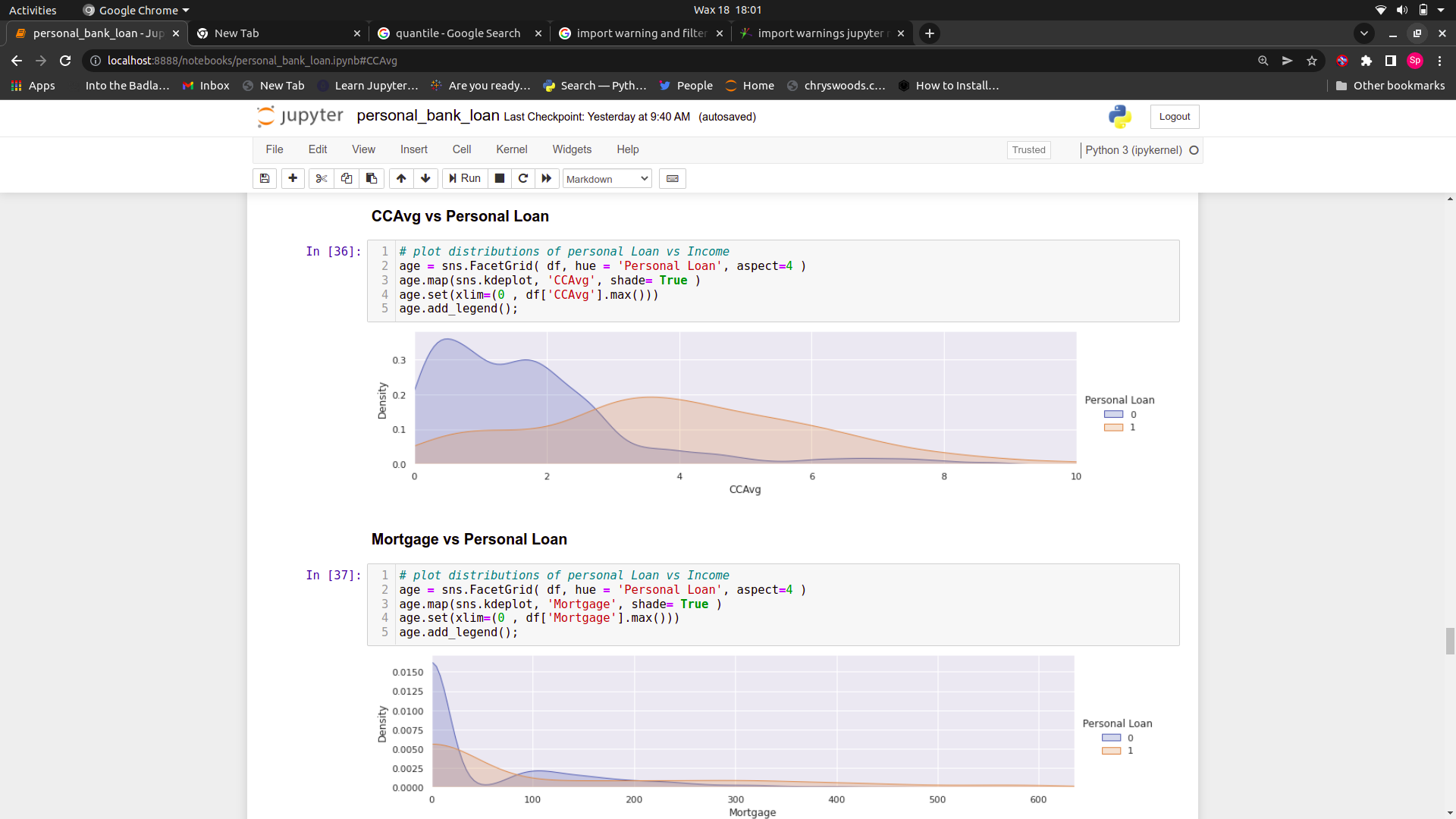
**Scatter-plot of Age vs. Experience**

****

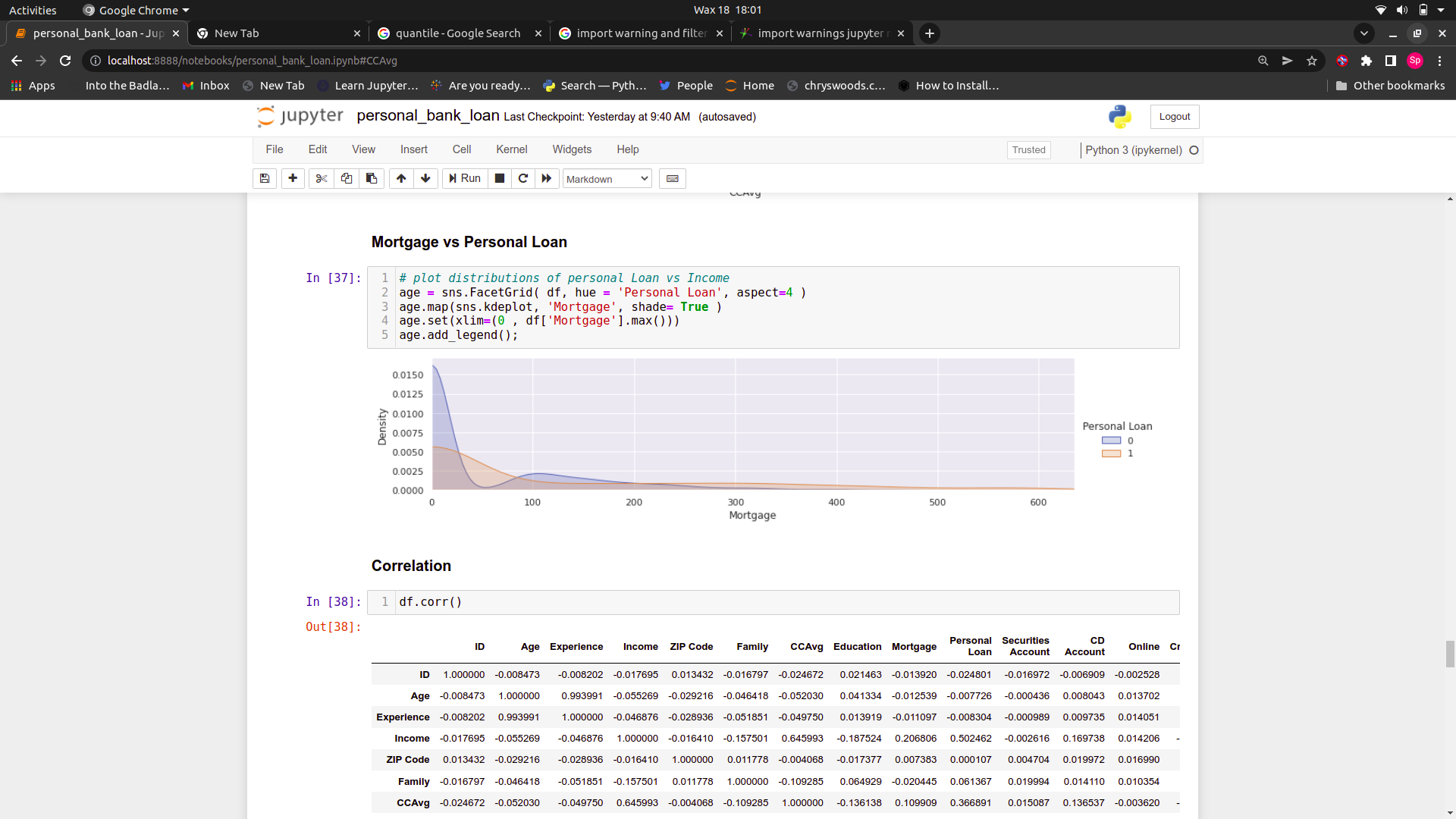
**Distribution of Income vs. Personal Loan**

****

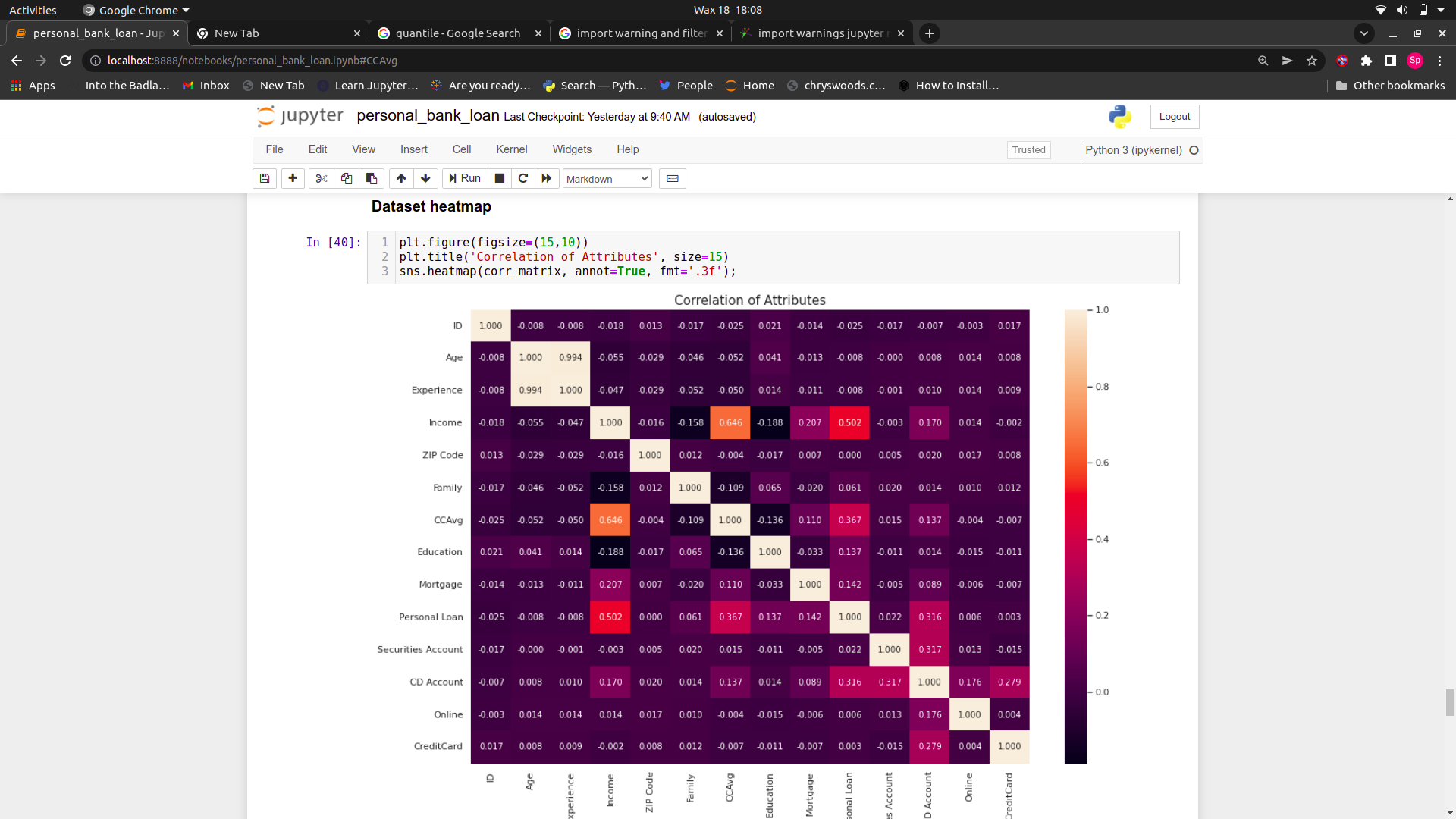
**Distribution of CCAvg vs. Personal Loan**

****

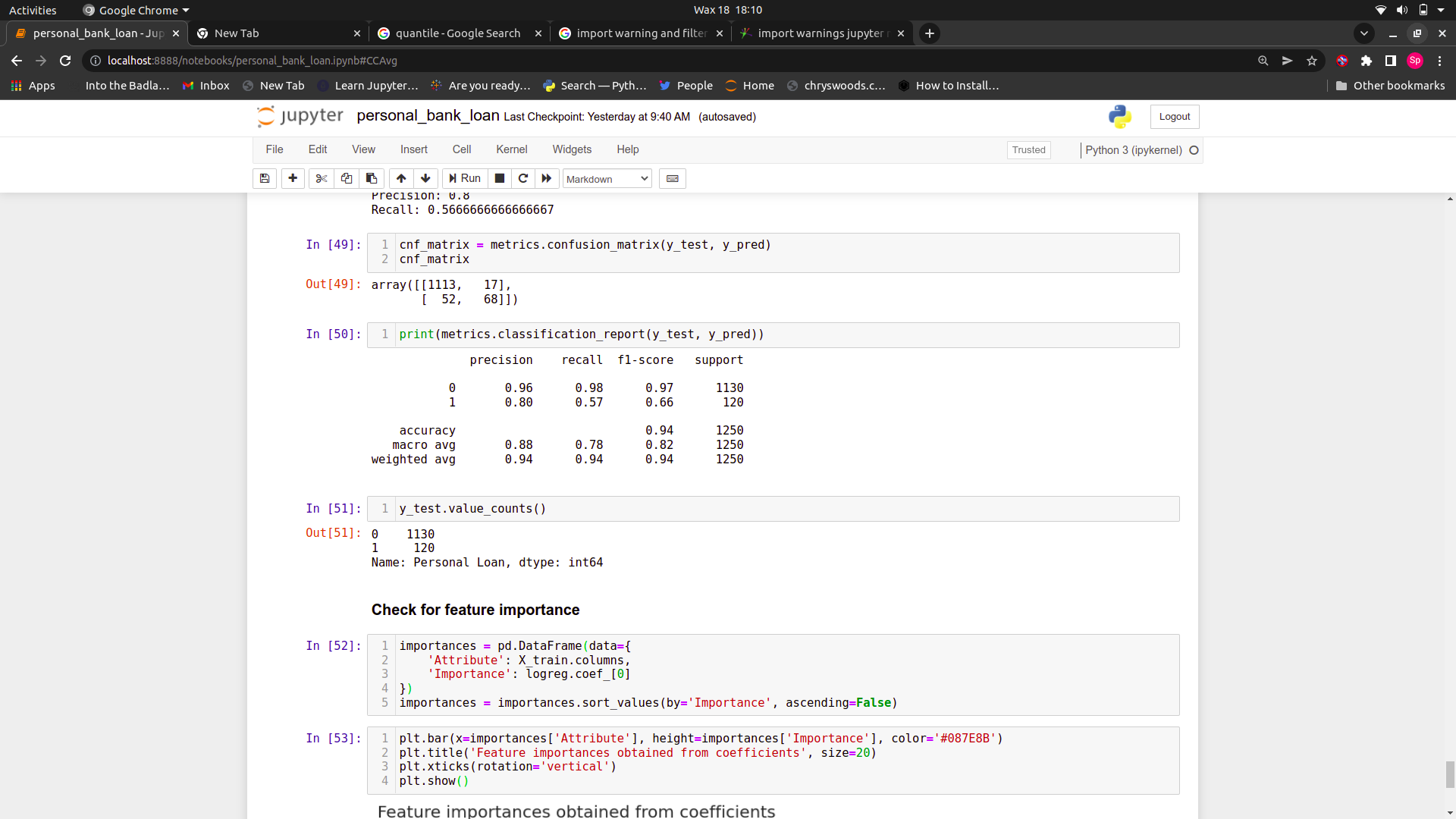
**Distribution of Mortgage vs. Personal Loan**

****

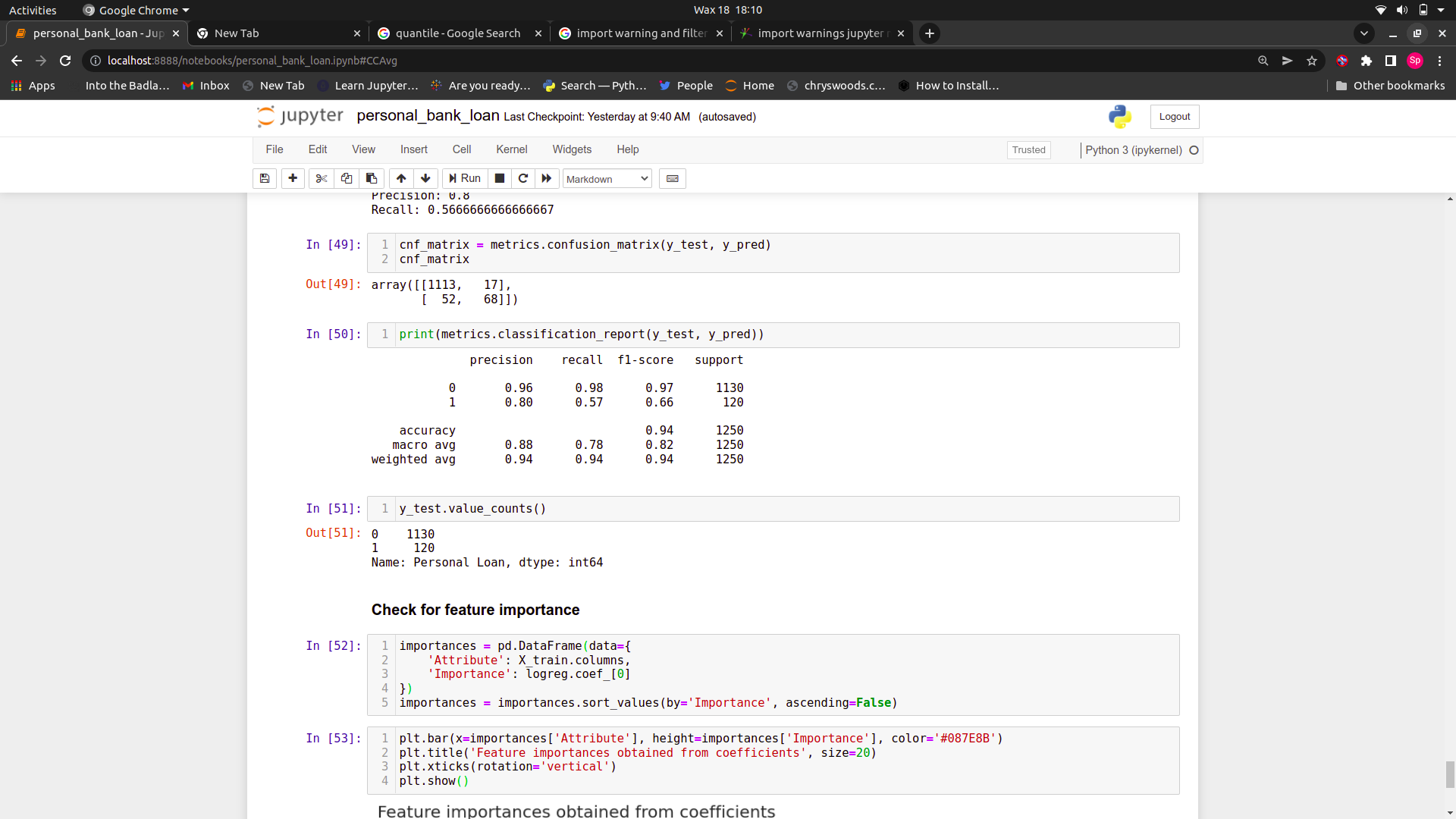
**HeatMap**

****

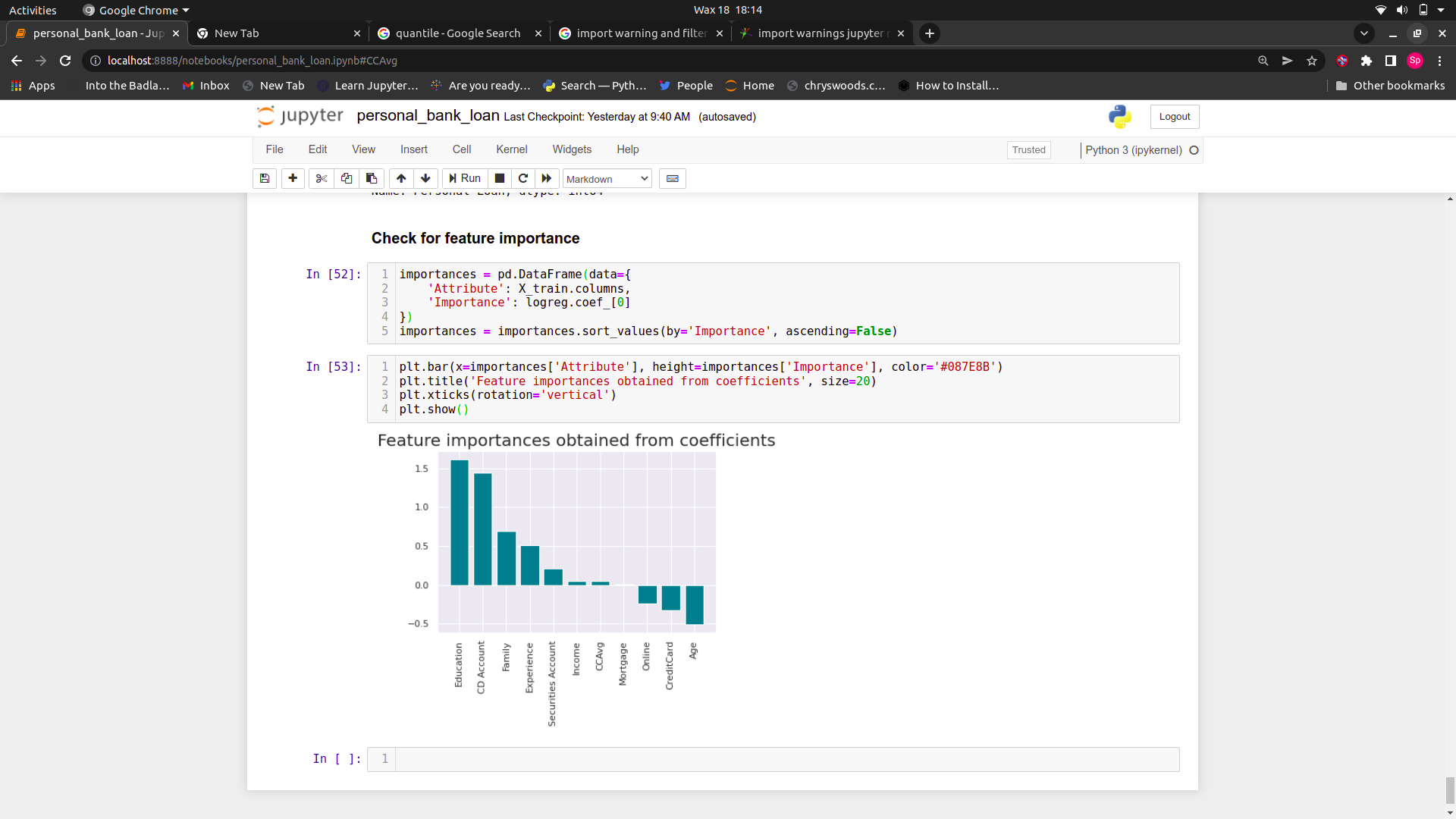
**Confusion Matrix**

****

**Classification Report**

****

**Feature Importance**

****