**Project Report**

**On**

**Customer churn prediction**

**Submitted as partial fulfillment for the award of**

**BACHELOR OF TECHNOLOGY**

**DEGREE**

**Session 2022-23**

**In**

**COMPUTER SCIENCE AND ENGINEERING**

**By**

**PANKAJ KUMAR**

**2000320100114**

**Under the guidance of**

**Dr Mala Saraswat**

**ABES ENGINEERING COLLEGE, GHAZIABAD**

|  |  |  |
| --- | --- | --- |
|  | APJAK logo | Image result for national board of accreditation logo |

**AFFILIATED TO**

**DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW**

**(Formerly UPTU)**

**STUDENT’S DECLARATION**

I hereby declare that the work being presented in this report entitled “Customer churn prediction” is an authentic record of my own work carried out under the supervision of Dr. Mala Saraswat”. The matter embodied in this report has not been submitted by me for the award of any other degree.

**Dated: 30/06/22** **Signature of student**

**(PANKAJ KUMAR)**

**Department: CSE**

## This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

|  |  |
| --- | --- |
|  | Signature of Supervisor(Dr Mala Saraswat)(Associate Professor)(CSE Department) |

## 

## CERTIFICATE

This is to certify that Project Report entitled “ Customer churn prediction ” which is submitted by ANKUR MISHRA in partial fulfillment of the requirement for the award of Data Science Project in Department of Computer Science Engineering of Dr. A.P.J. Abdul Kalam Technical University, formerly Uttar Pradesh Technical University is a record of the candidate own work carried out by him/them under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

**Supervisor MALA SARASWAT**

**Date 30/06/22**

ACKNOWLEDGEMENT

*It gives us a great sense of pleasure to present the report of the Data Science Project undertaken during B. Tech. Third Year. We owe special debt of gratitude to Dr. Mala Saraswat Department of Computer Science & Engineering, ABESEC Ghaziabad for his constant support and guidance throughout the course of our work. His/Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.*

*We also take the opportunity to acknowledge the contribution of Professor (Dr.) Divya Mishra, Head, Department of Computer Science & Engineering, ABESEC Ghaziabad for his full support and assistance during the development of the project.*

*We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.*

*Signature: PANKAJ KUMAR*

*Name : PANKAJ KUMAR*

*Roll No.:2000320100114*

*Date : 05/07/22*

ABSTRACT

In this project, we see how we can use machine-learning techniques and elementary data analysis to predict the customer churning. With a dataset of 7043 individuals containing features like ‘customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',

'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling','PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn', we attempt to predict the cutomer churn. We have used Decision Tree Algorithm for Analyzing the dataset.

Using data provided by www.kaggle.com, our goal is to apply machine-learning techniques to successfully predict which churned from the company. Features like OnlineSecurities, age, sex, and class will be used to make the predictions.

**TABLE OF CONTENTS**

page

DECLARATION ................................................................................................... ii

CERTIFICATE……....................................................................................................iii

ACKNOWLEDGEMENTS .................................................................................. iv

ABSTRACT ........................................................................................................... v

CHAPTER 1 7

1.1. ................................................................................................................. 7

1.2. ................................................................................................................. 8

CHAPTER 2 ……………………………. ......................................................... 9

2.1. ............................................................................................................... 9

2.1.1. ....................................................................................................... 10

2.1.2. ....................................................................................................... 10

CHAPTER 3 ………………………….….......................................................... 11

4.1. ................................................................................................................11

4.2. ................................................................................................................12

CHAPTER 4 (CONCLUSIONS) ......................................................................13

REFERENCES... .................................................................................................... 15

**CHAPTER 1**

**INTRODUCTION**

The goal of the project is to the customers who have churned based off a set of data,

Which customers are churning? Why are they churning?

I have used Kaggle competition "WA\_Fn-UseC\_-Telco-Customer-Churn" (see <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>) to retrieve necessary data and evaluate accuracy of our predictions. The historical data has been split into two groups, a 'training set' and a 'test set'. For the training set, we are provided with the outcome (whether or not a customer churn). We used this set to build our model to generate predictions for the test set.

Predict behavior to retain customers. we can analyze all relevant customer data and develop focused customer retention programs. [IBM Sample Data Sets].

• Programming language Python and its libraries NumPy (to perform matrix operations) and SciKit-Learn (to apply machine learning algorithms)

• Machine Learning Algorithm Decision Tree.

• Feature Engineering techniques

Each row represents a customer, each column contains customer’s attributes described on the column Metadata.

The data set includes information about:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents

1.1Problem Introduction

* + 1. Motivation
    2. Project Objective:

Churn analysis is the evaluation of a company’s customer loss rate in order to reduce it. Also referred to as customer attrition rate, churn can be minimized by assessing your product and how people use it.. The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of customers churned. After this the results of applying machine learning models are compared and analyzed on the basis of accuracy.

* + 1. Scope of the Project

This project involves implementation of data analytics and machine learning. This project work can be used as reference to learn implementation of machine learning from very basic. In future the idea can be extended by making more advanced graphical user interface with the help of newer libraries like shiny in R. An interactive page can be made, i.e. if the value of a attribute is changed on the scale the values corresponding to its graph (ggplot or histogram) will also change. We can also draw much focused conclusions by combining results we obtained.

For now many new bussiness are coming to the market so holding customers for a company/organization or a shop becomes important factor to survive for that company. So this model can be used to analyse where are they lacking and why their customers are churned and what changes they can make in their rules or maintain their products according to the analysis, so that they can hold their existing customers and can add more.

**CHAPTER 2**

**LITERATURE SURVEY**

Feature engineering is the most important part of data analytics process. It deals with, selecting the features that are used in training and making predictions. In feature engineering the domain knowledge is used to find features in the dataset which are helpful in building machine learning model. It helps in understanding the dataset in terms of modeling. A bad feature selection may lead to less accurate or poor predictive model. The accuracy and the predictive power depend on the choice of correct features. It filters out all the unused or redundant features.

Based on the exploratory analysis above, following features are used age, sex, cabin, title, Pclass, family size (parch plus sibsp columns), fare, embarked. CustomerID is chosen as response column. These features are selected because their values have an impact on the rate customer churn.

2.1 **Machine Learning Models**

Various machine learning models are implemented to validate

and predict survival.

2.1.1 **Logistic Regression**

Logistic regression is the technique which works best when

dependent variable is dichotomous (binary or categorical).

[23] The data description and explaining the relationship

between one dependent binary variable and one or more

nominal, ordinal, interval or ratio-level independent variables

is done with the help of logistic regression. It is used to solve

binary classification problem, some of the real life examples

are spam detection- predicting if an email is spam or not,

health-Predicting if a given mass of tissue is benign or

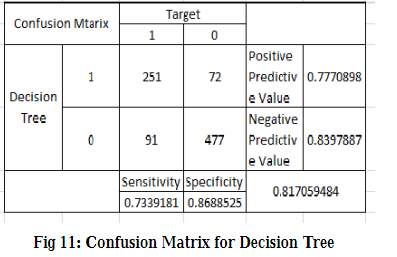
malignant, marketing- predicting if a given user will buy an

insurance product or not.

2.1.2 **Decision Tree**

Decision tree is a supervised learning algorithm. This is Generally used in problems based on classification. It is suitable for both categorical and continuous input and output variables. Each root node represents a single input variable (x) and a split point on that variable. The dependent variable (y) is present at leaf nodes. For example: Suppose there are two independent variables, i.e. input variables (x) which are height In centimeter and weight in kilograms and the task to find Gender of person based on the given data. (Hypothetical Example, for demonstration purpose only)

will be the value of “x” in the bar-plots. If wrong features where selected then even the good algorithm may produce the bad predictions. Therefore, feature engineering acts like a backbone in building an accurate predictive model.



will be the value of “x” in the bar-plots. If wrong features where selected then even the good algorithm may produce the bad predictions. Therefore, feature engineering acts like a backbone in building an accurate predictive model

**CHAPTER 3**

# IMPLEMENTATION AND RESULTS

* 1. **Software and Hardware Requirements**

**Memory**

8GB

**CPU**

Intel i3 minimum

Intel i5, i7 or i9 preferred

(M1 NOT recommended)

**Hard Drive**

SSD is preferred  
500GB minimum

(or 256 GB and an external hard drive)

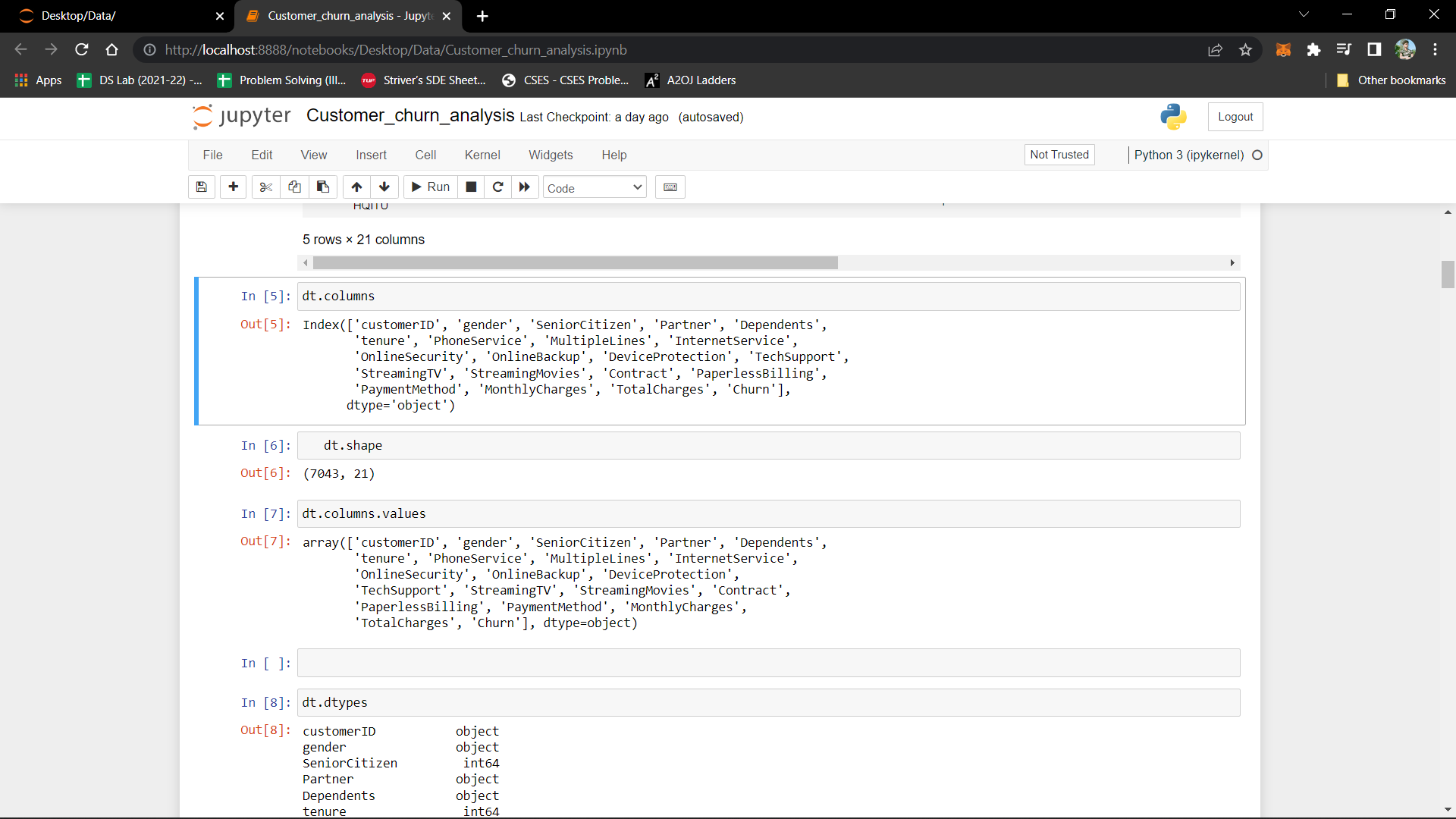
**Operating Systems**

Windows 10 (Home or Pro)

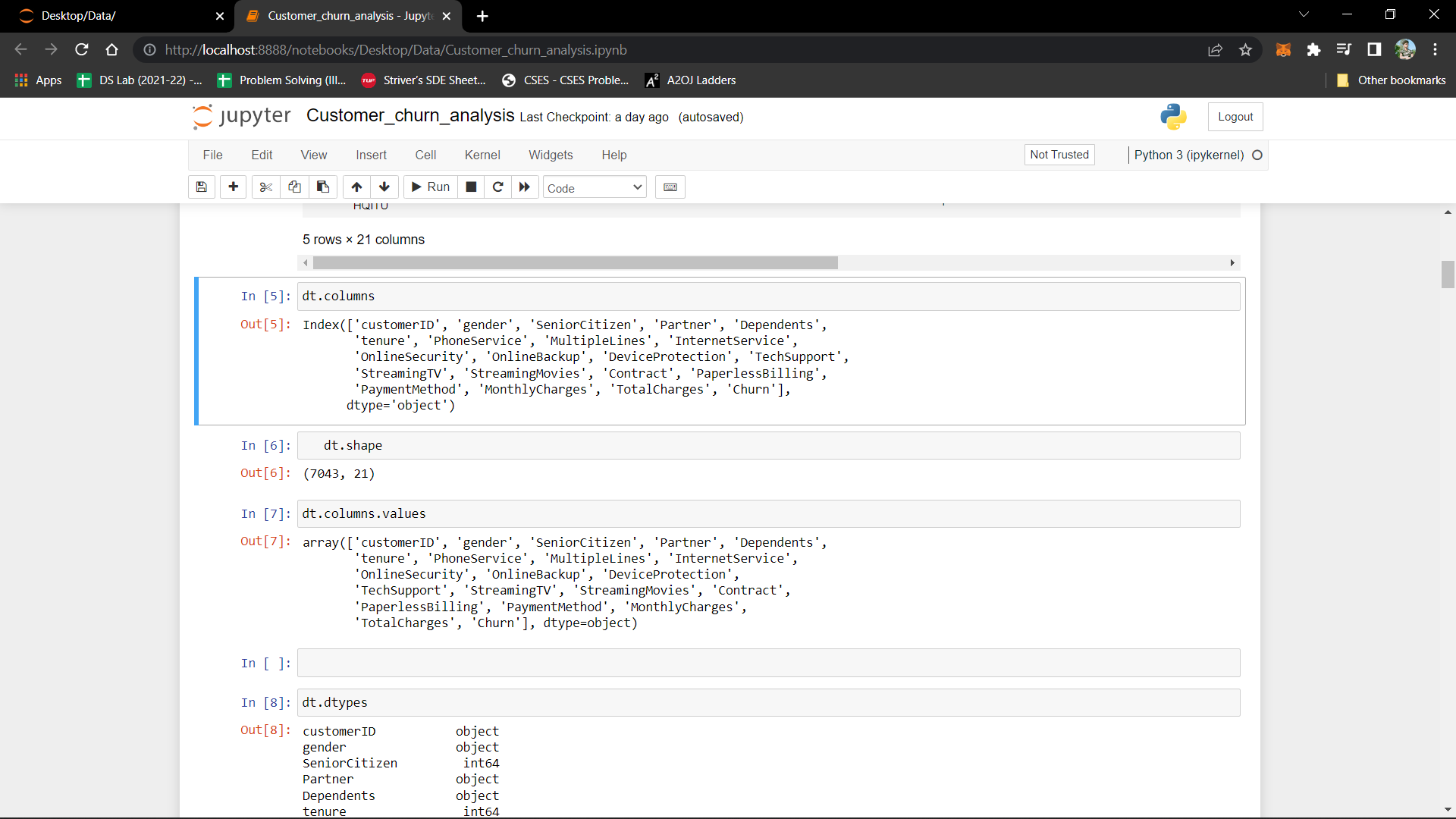
Mac: High Sierra, Mojave, or Catalina

Linux: recent distribution

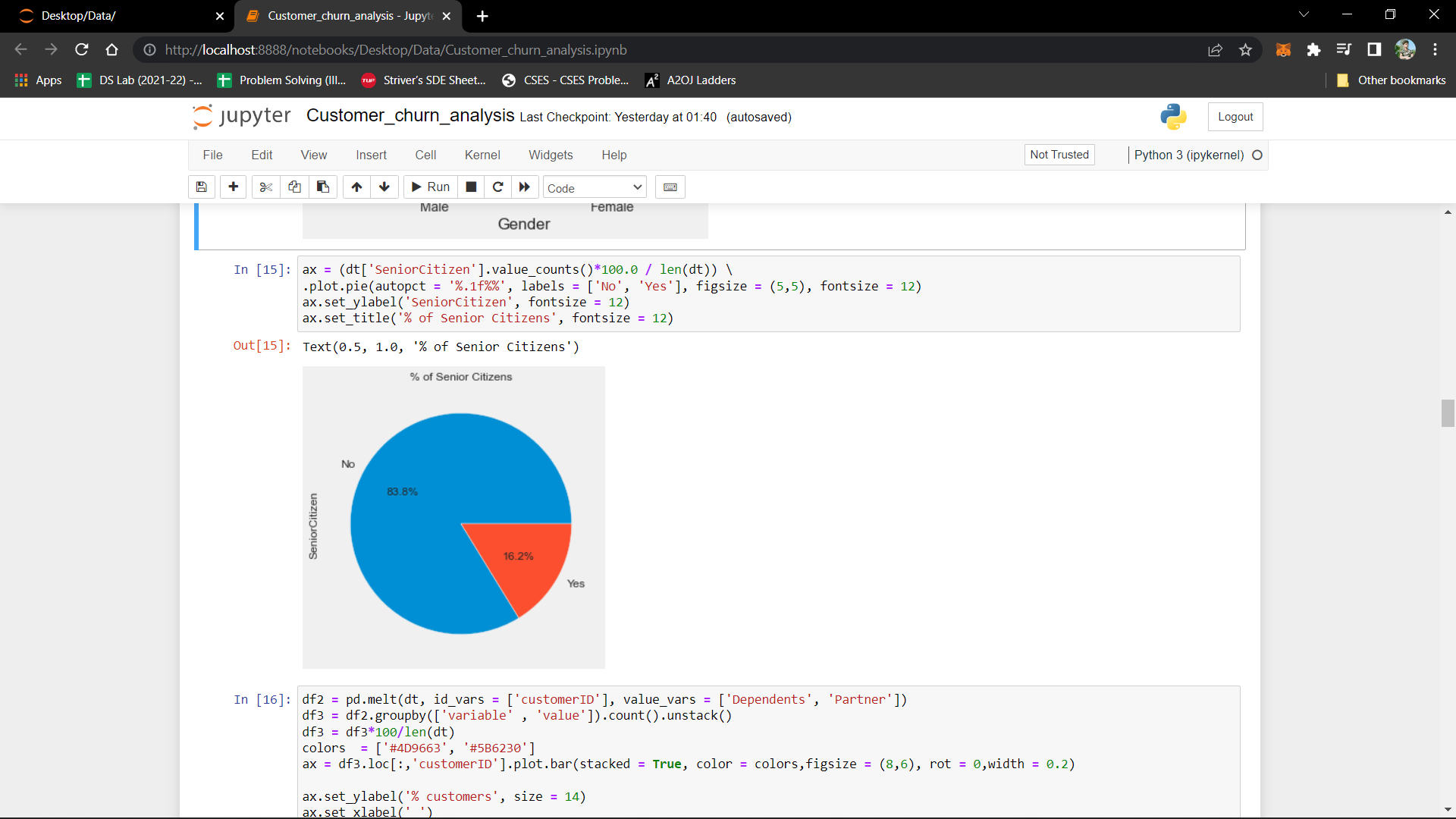
* 1. **Implementation Details**
     1. **Snapshots Of Interface**
* These are the columns of dataset which contain different values for all Customers.

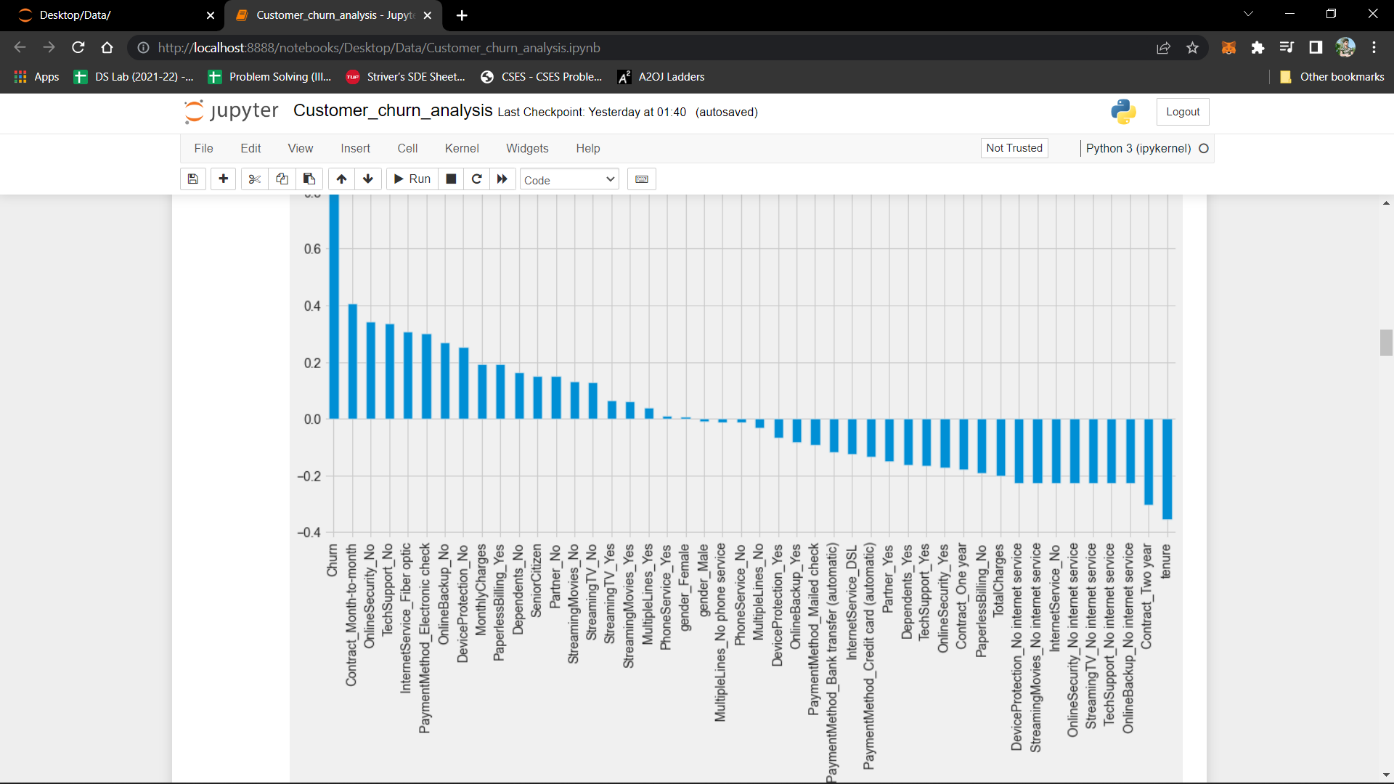
****

* These are the no. of rows and columns of dataset.

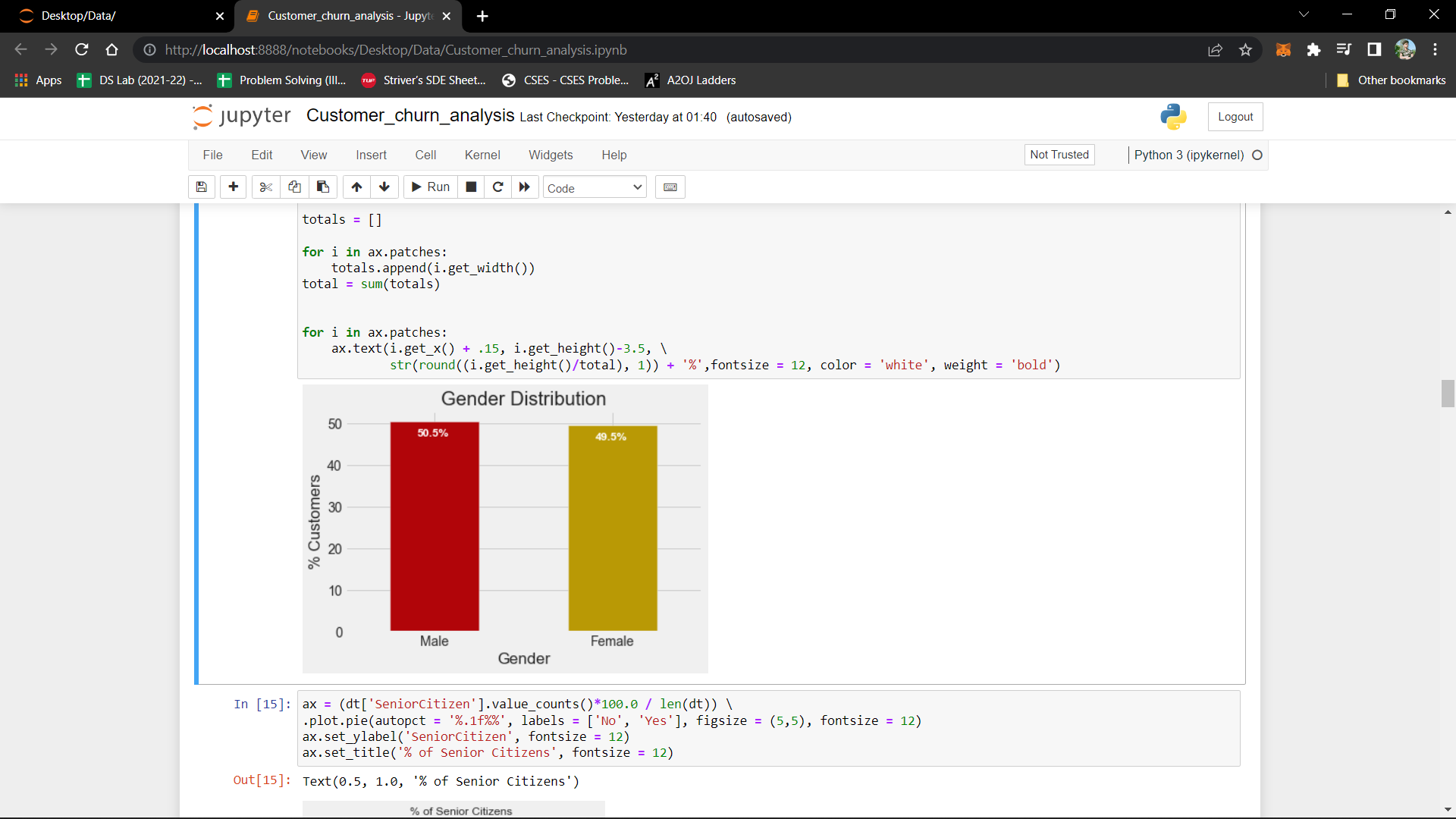
****

* **Percentage of senior citizen.**

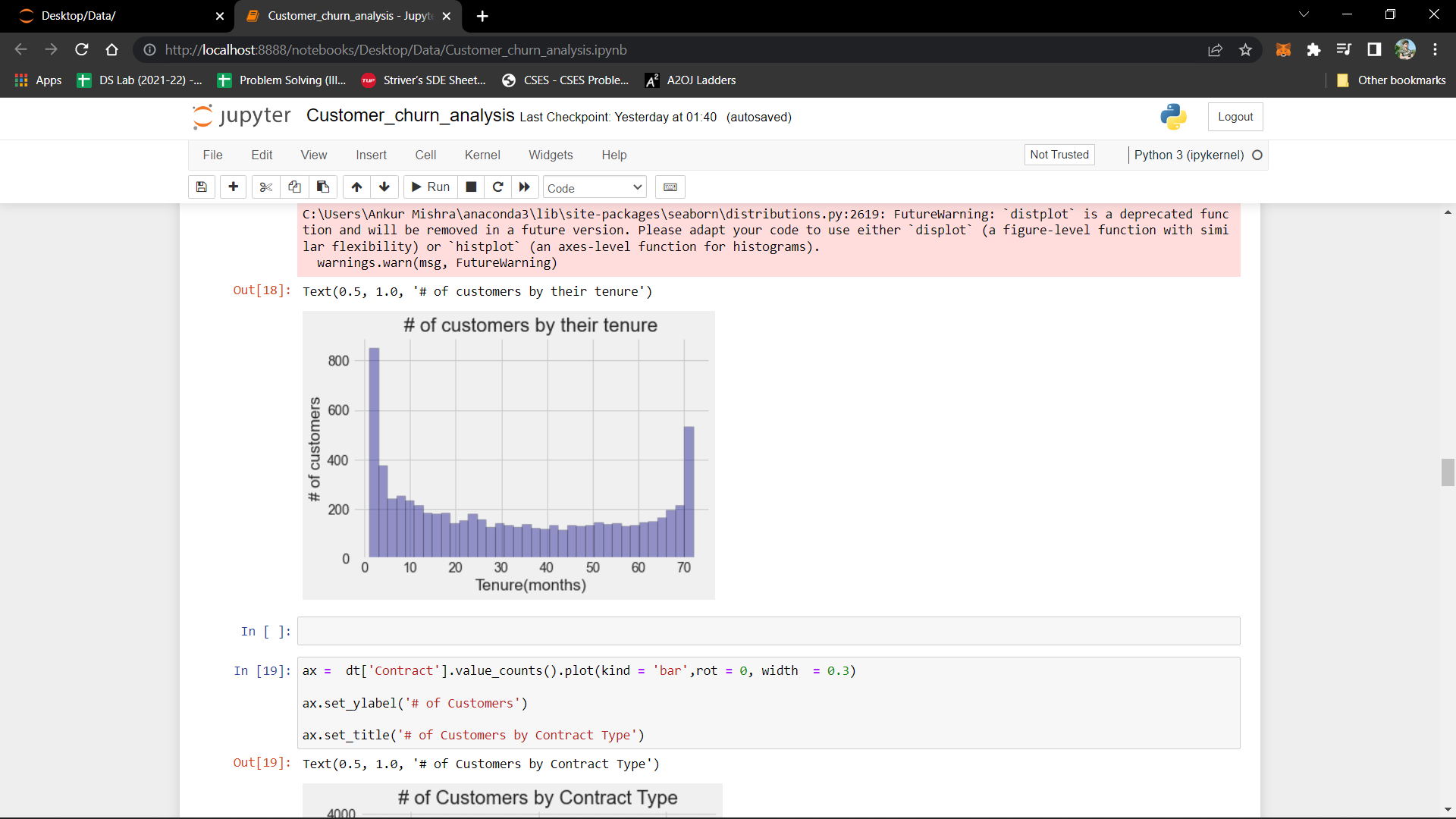
****

* **All the columns by their numbers.**

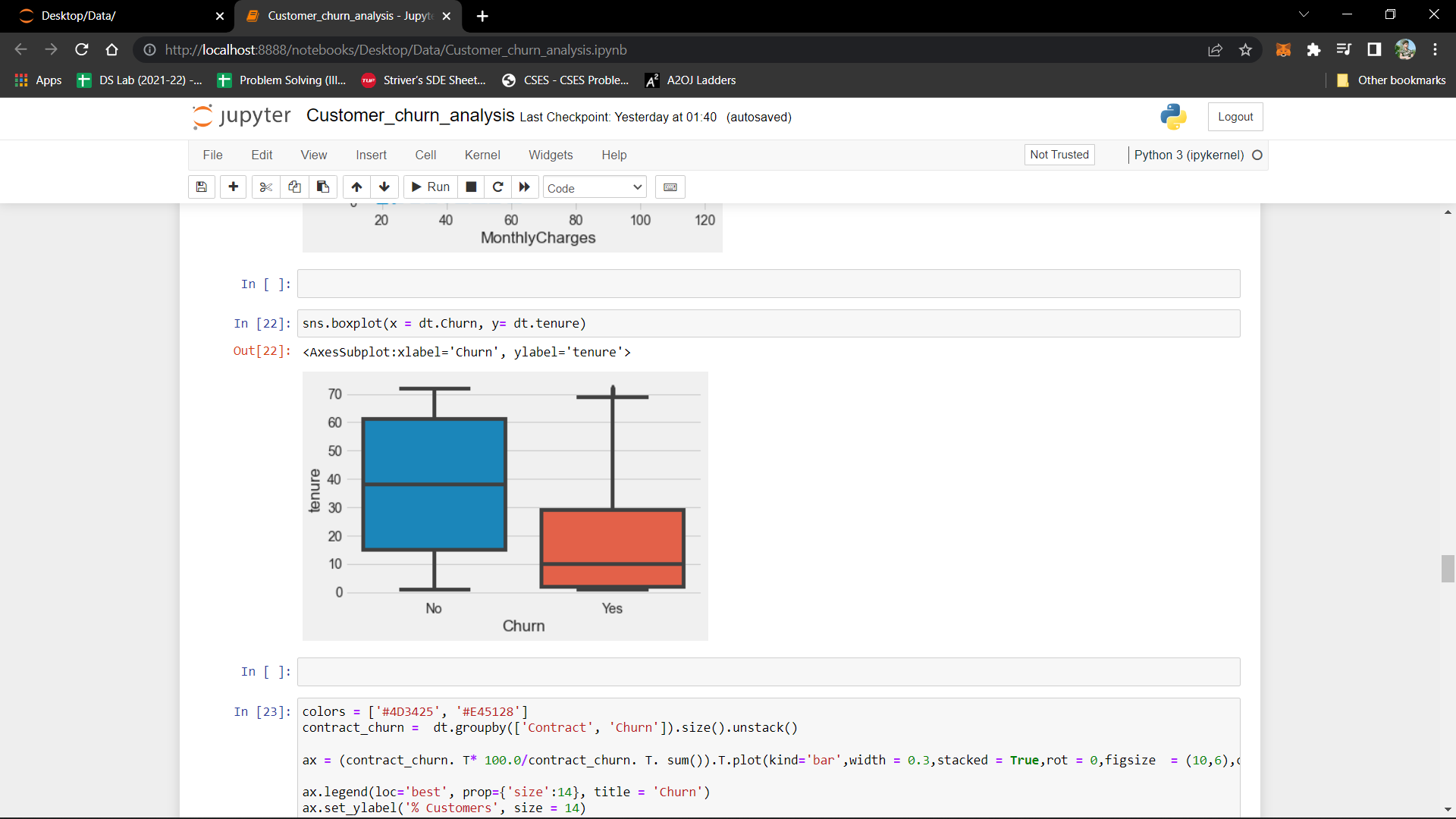
* **Gender distribution plot.**



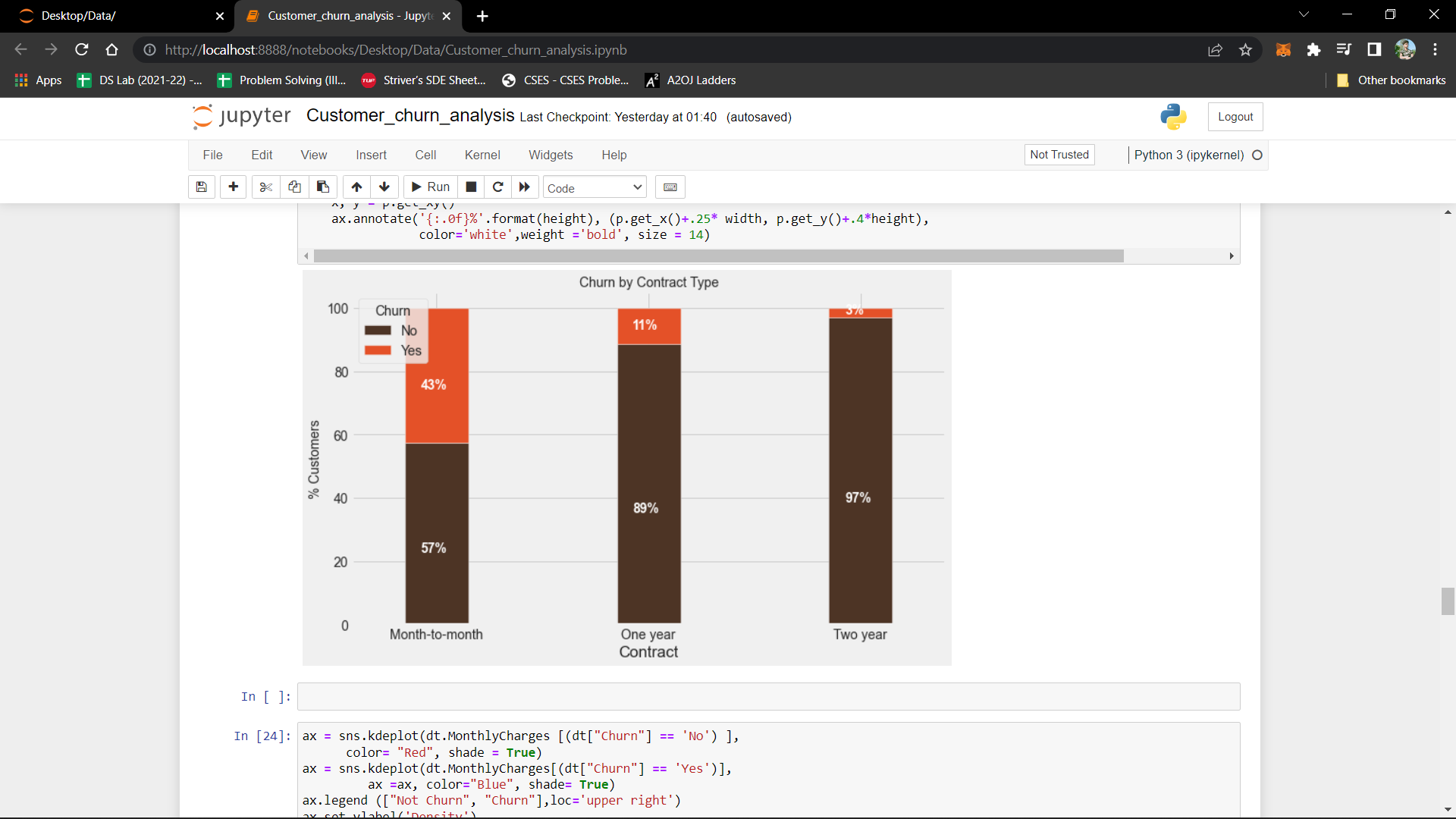
* **Customers by their tenure**

****

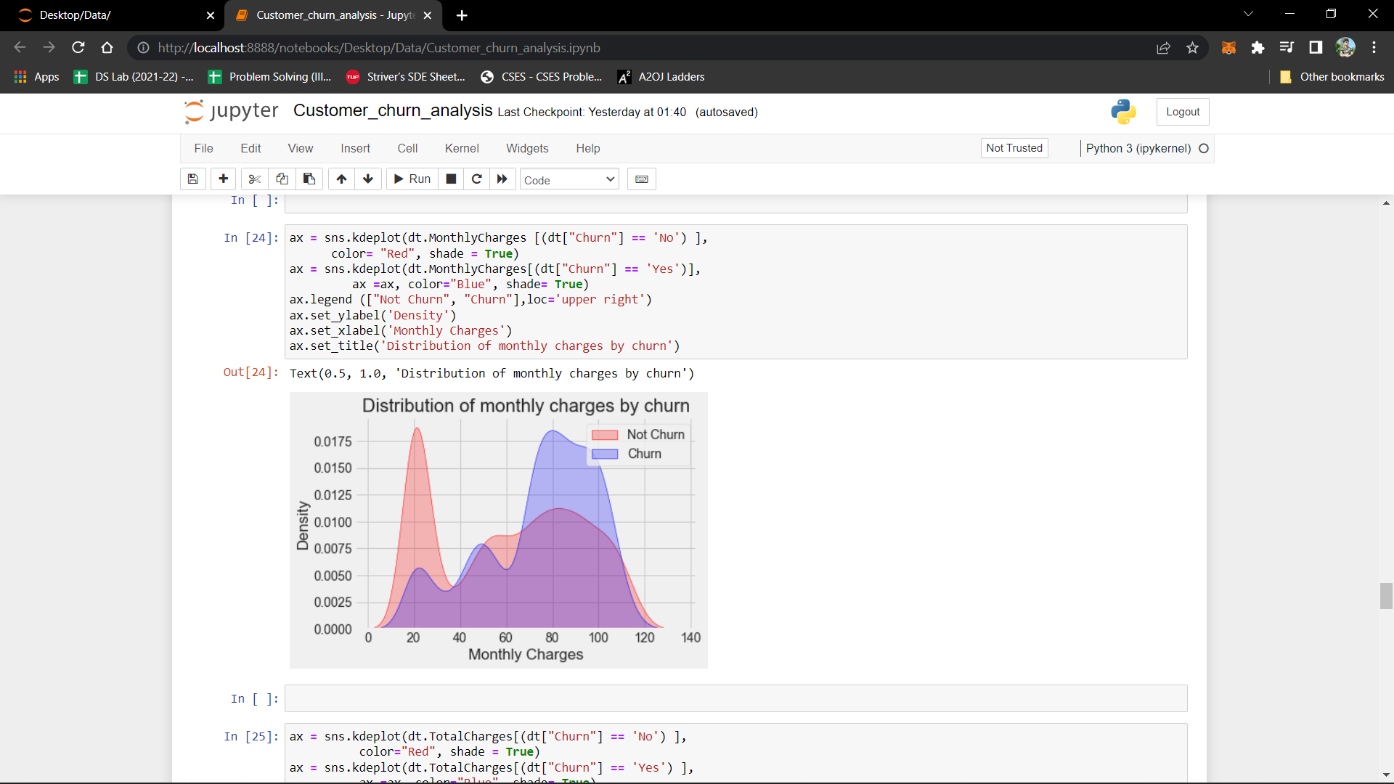
* **Tenure of churn customers**

****

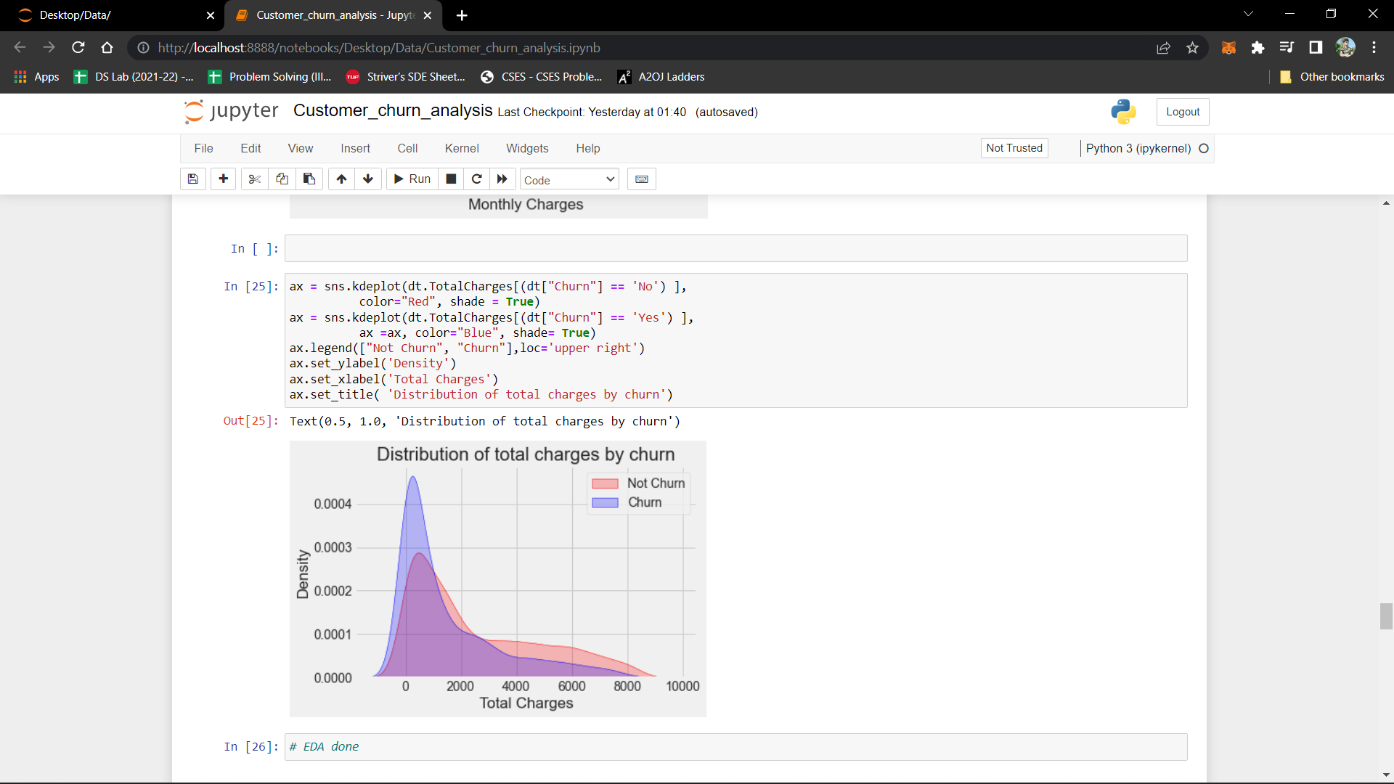
* **Churn by their contract time**

****

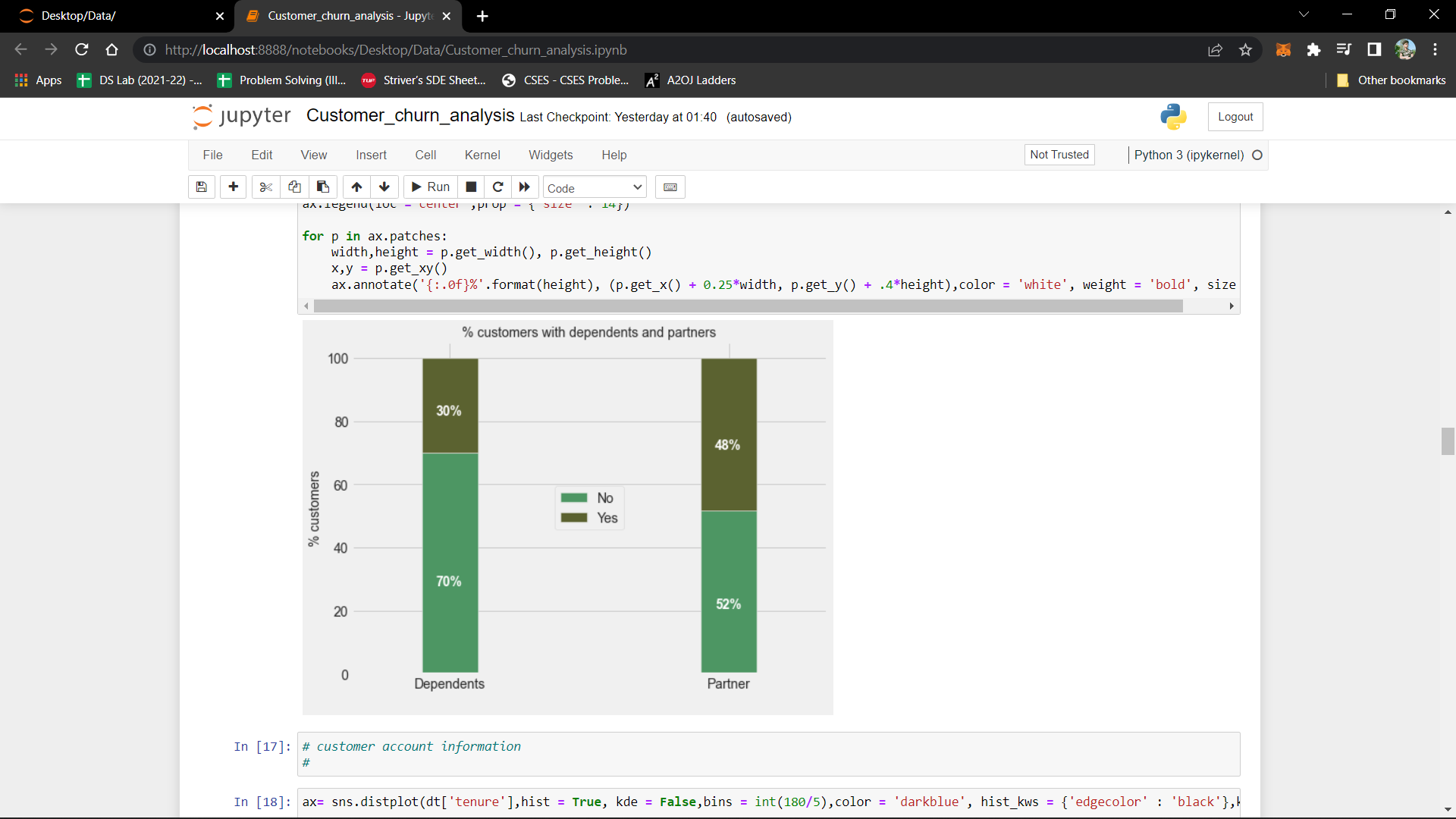
* **Distribution of monthly charges by churn Vs Monthly charges**

****

* **Distribution of monthly charges by churn Vs Total charges**

****

* **Percentage of customers who depends with partners**

****

**Test Cases**

Train.csv

Test.csv

* + 1. **Results**

The given dataset is implemented using Decision Tree.

**CHAPTER 4**

# CONCLUSION

* 1. Performance Evaluation

As a result of our work, we gained valuable experience of building prediction systems and achieved our best score on Kaggle: 80.383% of correct predictions (in Kaggle leader board, it corresponds to positions 477 - 881 out of 3911 customers).

• We performed featured engineering techniques

• Changed alphabetic values to numeric

• Calculated family size

• Used linear regression algorithm to fill in missing ages

• We used prediction algorithm in python

• Decision tree

• We achieved our best score 80.383% correct predictions

* 1. Future Directions

Of course there is still room for improvement, like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features. Another thing that can improve the overall result would be a more extensive hyper-parameter tuning on several machine learning models. You could also do some ensemble learning.

**References**

[1] dataset from Kaggle

<https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

[2] You tube channel

<https://www.youtube.com/watch?v=VuiKM1Yol1U>

[3] studying details about churn analysis

<https://baremetrics.com/blog/churn-analysis>