**BIRLA INSTITUTE OF TECHNOLOGY**

**MESRA**



**MASTER OF COMPUTER APPLICATIONS (SEM - IV)**

**2020-2021**

**Machine Learning Project**

**(Documentation of the Assign Project)**

**SUBMITTED TO – SUBMITTED BY –**

**Prof.RASHMI RATHI UPADHYAY KESHAV TYAGI**

**(Assistant Professor) (ROLL – MCA/10054/19)**

# **Hotel-Booking-Cancellation-Prediction Project**

**Abstract**

Booking cancellations negatively contribute to the production of accurate forecasts, which comprise a critical tool in the hospitality industry. Research has shown that with today’s computational power and advanced machine learning algorithms it is possible to build models to predict bookings cancellation likelihood. However, the effectiveness of these models has never been evaluated in a real environment. To fill this gap and investigate how these models can be implemented in a decision support system and its impact on demand-management decisions, a prototype was built and deployed in two hotels. The prototype, based on an automated machine learning system designed to learn continuously, lead to two important research contributions.

In reservation-based industries, an accurate booking cancellation forecast is of foremost importance to estimate demand. By combining data science tools and capabilities with human judgement and interpretation, this aims to demonstrate how the semiautomatic analysis of the contribute to synthesizing research findings and identify research topics about booking cancellation forecasting.

Furthermore, The data used was obtained through a keyword search in Scopus and Web of Science databases. The methodology presented not only diminishes human bias, but also enhances the fact that data visualisation and text mining techniques facilitate abstraction, expedite analysis, and contribute to the improvement of reviews. Results show that despite the importance of bookings’ cancellation forecast in terms of understanding net demand, improving cancellation, and overbooking policies, further research on the subject is still needed.

**Project introduction:**

The cancellation rate for booking hotels online is high that creates discomfort for many hotels and create a desire to take precautions. Therefore, predicting reservations that can be cancelled will create a surplus value for hotels and hotels can take action to prevent these cancellations. In my final project, I will try to explore the dataset and explain how to predict future cancelled reservations in advance by machine learning methods.

**What we are going to do in this project:**

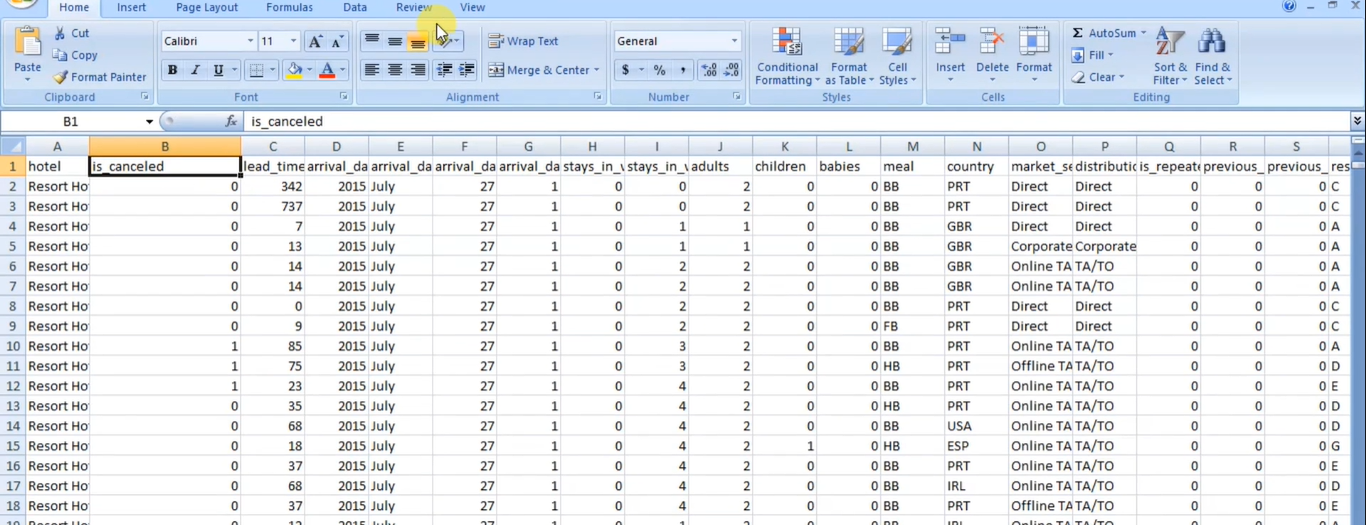
We are using **(Google Colab + Juputer Notebook)** during the develpoment of this Project.

Advantages of using **Google Colab :**

1. Like Jupyter Notebook + Bonus feature.
2. Hosted By Google.
3. Required Google account.
4. No Installation required.
5. Network speed is very fast.
6. Write Python code directly inside browser.
7. Notebook File stored in Google Drive.
8. CPU,GPU and TPU Support.

Again, I am clarifying the problem statement that we a making such a machine learning model that can predict wheather a particular booking which has been done by a user is going to cancel or not.

This is exactly our that **feature** that we have to predict considering all data set.



You can see how much huge chunk of data we have over here and considering is all these features, we have to build such a machine learning that can predict wheather a particular booking is going to cancel or not.

**But before building such a model**, you have to understand your data or what your data is all about.

So the best way to understand new data is that performing lots of analysis or new data by fact and some amazing insight from this huge chunk of data.

And once we understand about data,what my data is all about, what exactly did i think once we understand our data to a greater extent than we are going to build such a model that can predict what exactly is….

It is back to this future by doing lots of data processing, doing lots of feature encoding techniques, changing techniques ,dealing with missing value and lots of machine learning algorithns, we are going to apply on our data.

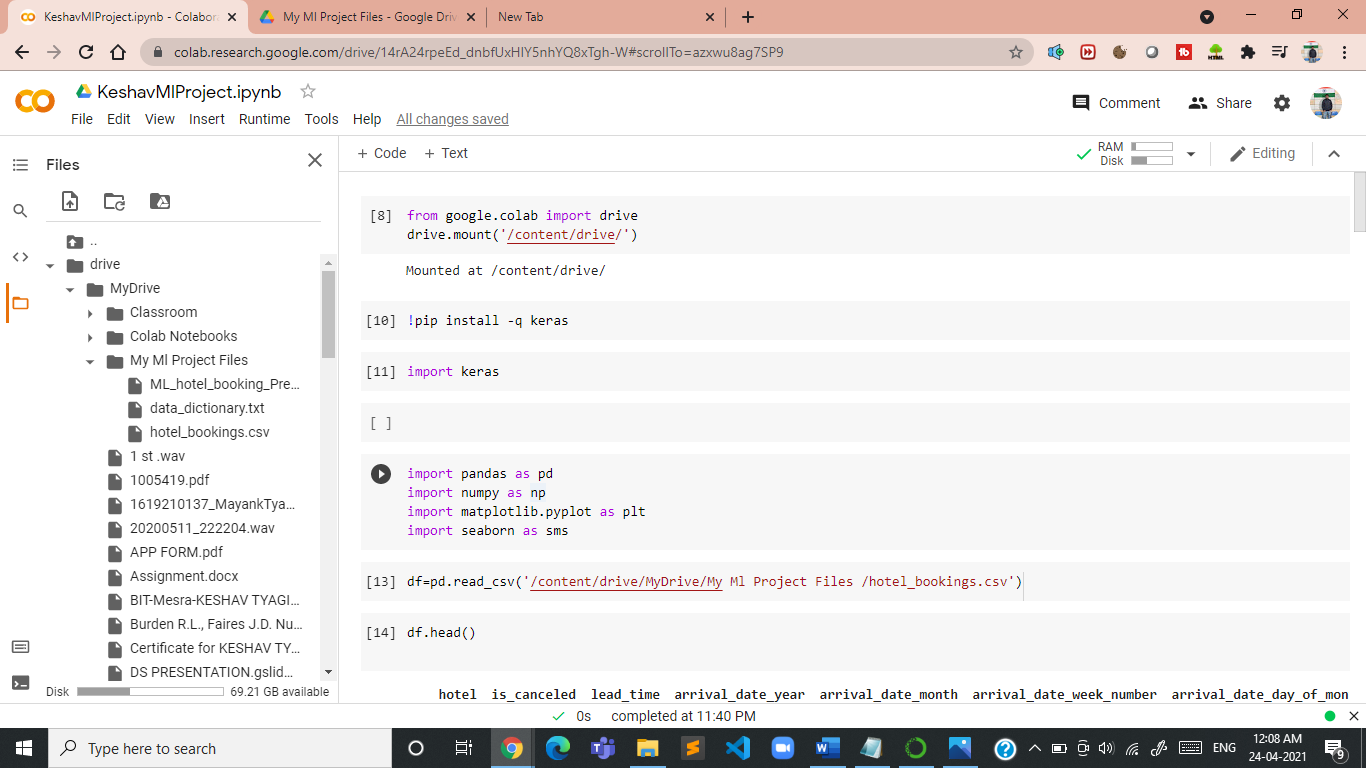
**1.Perform Data Cleaning & Prepare your Data for Modelling Purpose.**

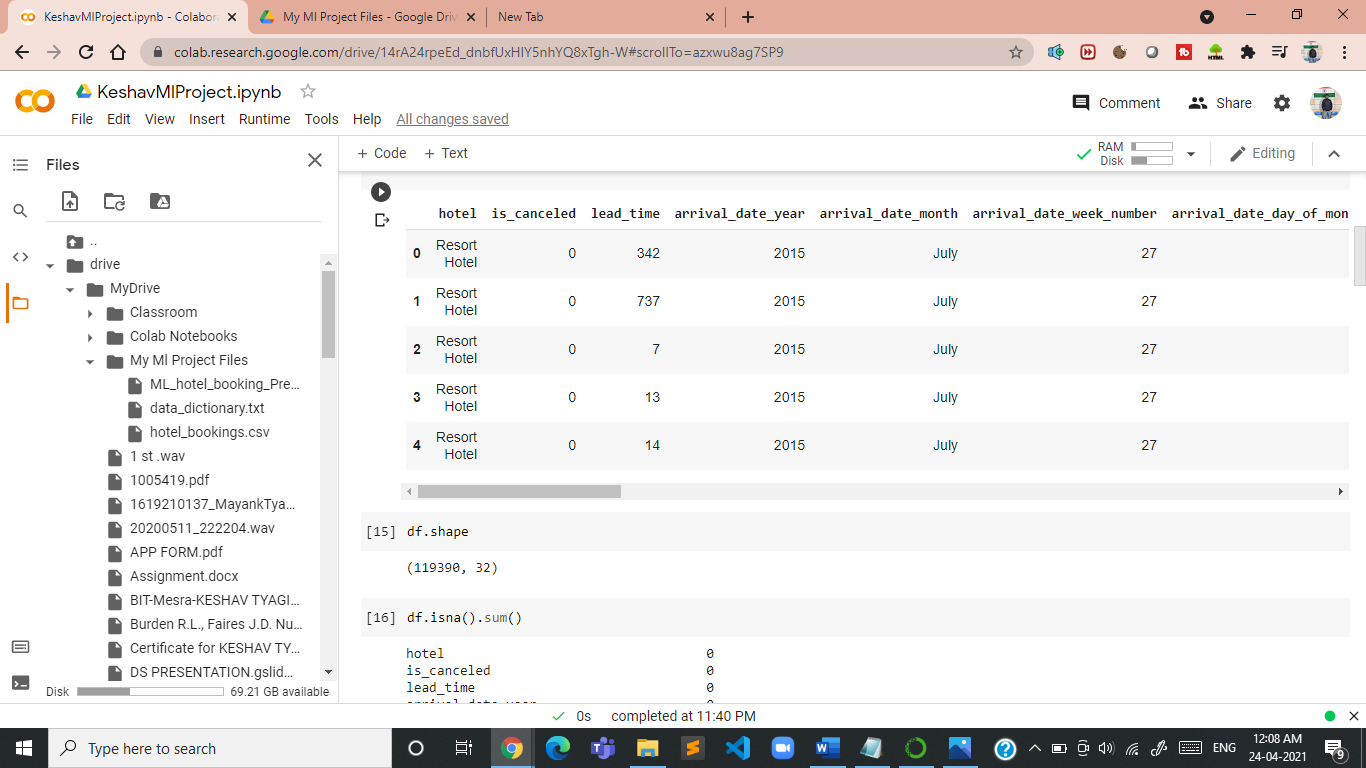
**# Replace missing values:**

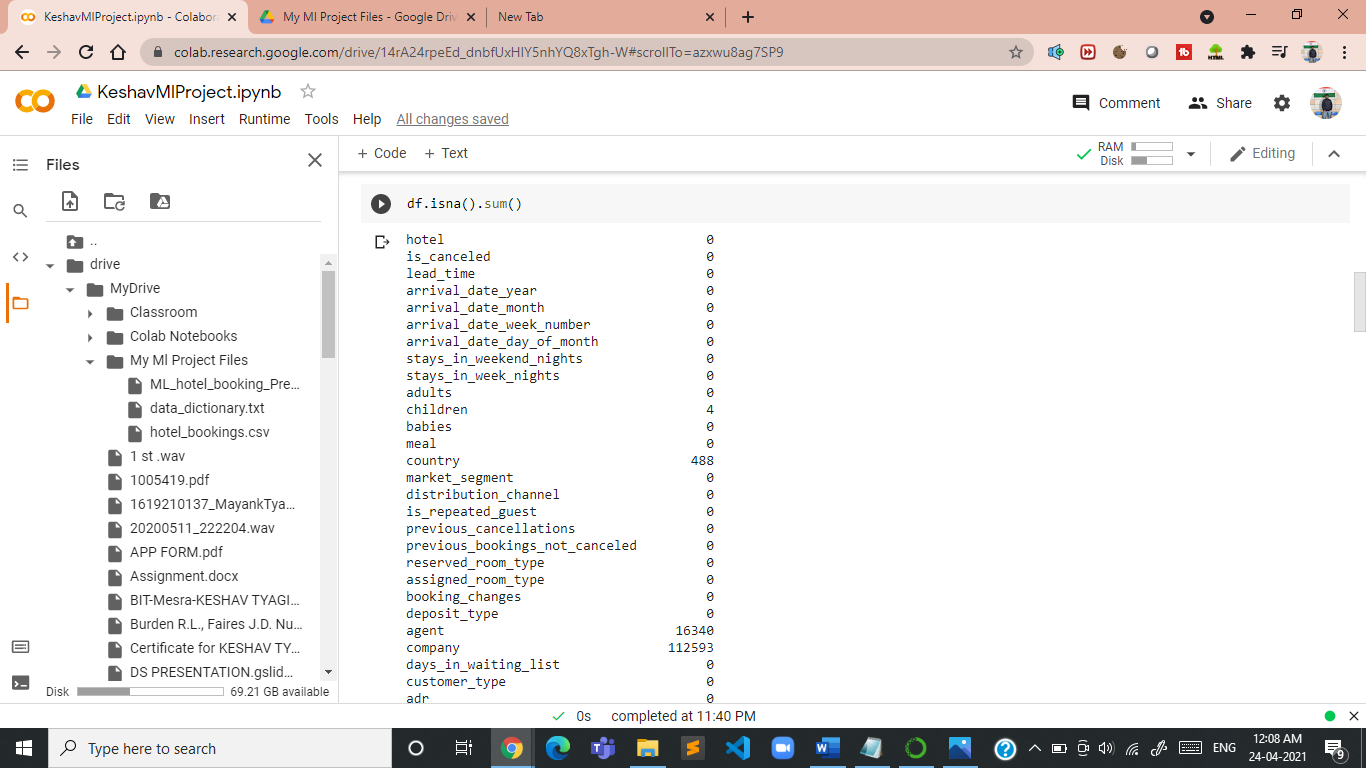
**# agent: If no agency is given, booking was most likely  made without one.**

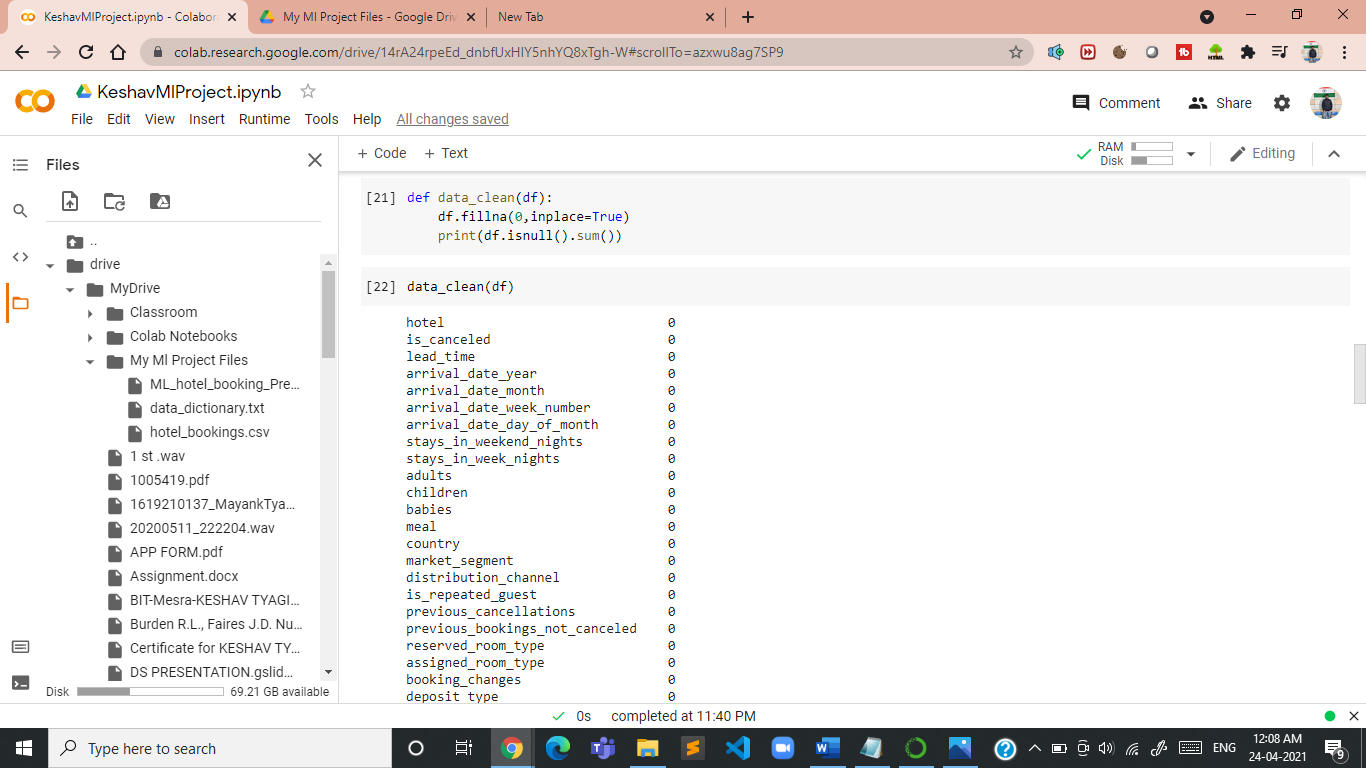
**# company: If none given, it was most likely private.**

**# rest should be self-explanatory.**





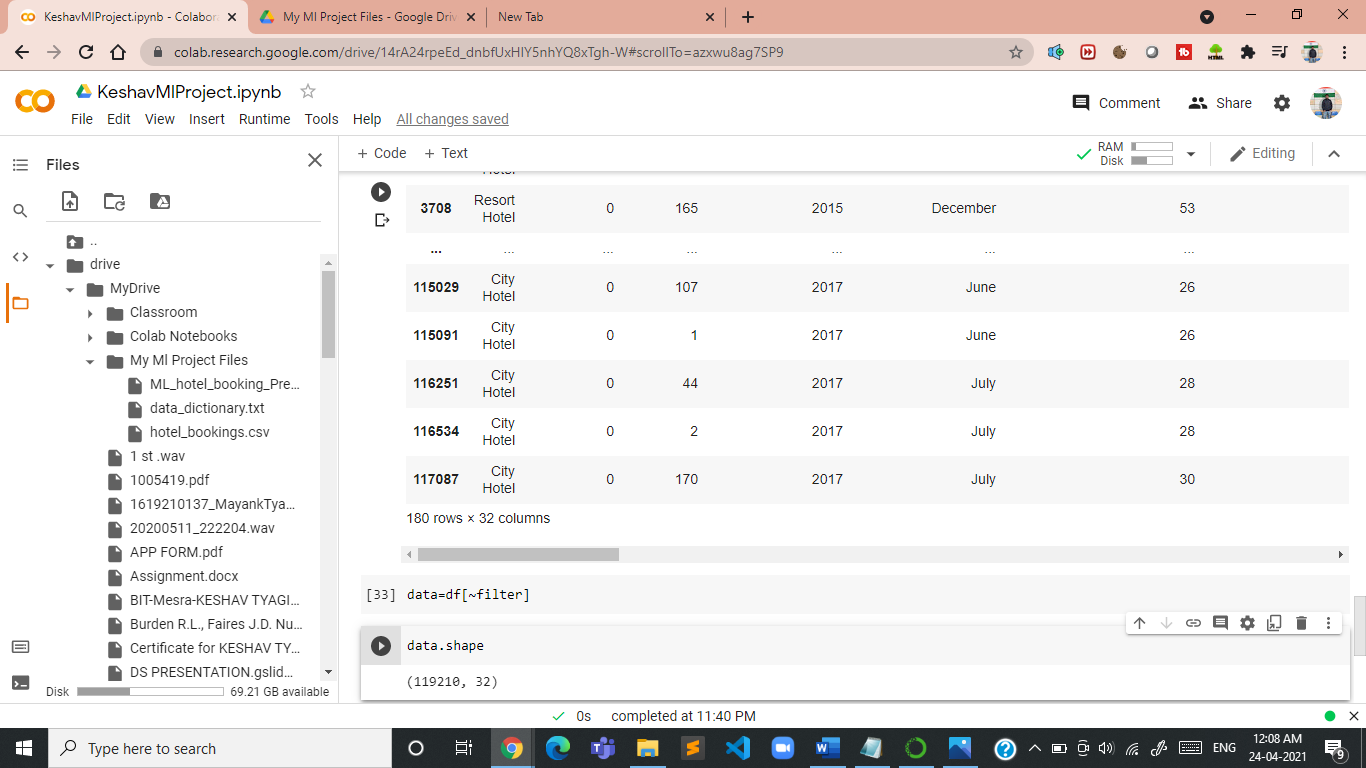


****

**### seems to have some dirtiness in data as Adults,babi es & children cant be zero at a same time.**

**### Visualise Entire Dataframe where adult,children & babies are 0**

**This is exactly that data on which you have to perform all your analysis on which you have to build your machine learning model after doing lots of feature engineering on new data.**

****

## Now we find where do the guests come from?Lets perform Spatial Analysis

## So spatial Analysis is all about whenever you are going to visualize the data on some map so that you get a clear cut yapped from which which location your guests are coming.

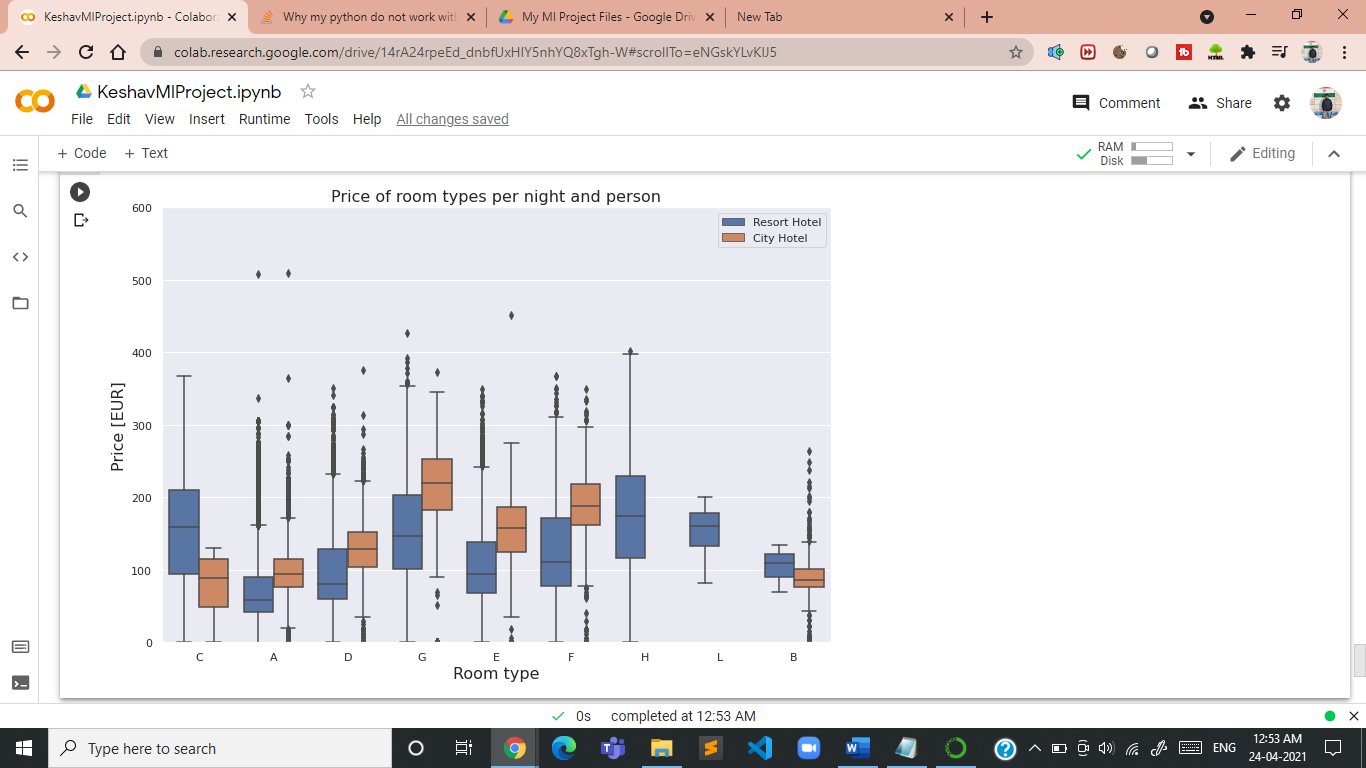
## 

So, that’s type of entrance ,how you can fetch stronger data.

#### **People from all over the world are staying in these two hotels. Most guests are from Portugal and other countries in Europe**

## Now we find how much the guests pay for a night ?Lets use distribution function or of a plot,or you can consider some fancy stuff like box plot.

#### Both hotels have different room types and different meal arrangements. Seasonal factors are also important. So the prices vary a lot. Since no currency information is given, but Portugal is part of the European Monetary Union, I assume that all prices are in EUR.



After all you get this visual with respect to each of your room type over here. You will see this blue line exactly a resort hotel and this orange one with middle line exactly your median.

#### **This figure shows the average price per room, depending on its type and the standard deviation. Note that due to data anonymization rooms with the same type letter may not necessarily be the same across hotels.**

## Now find how does the price per night vary over the year?

## 

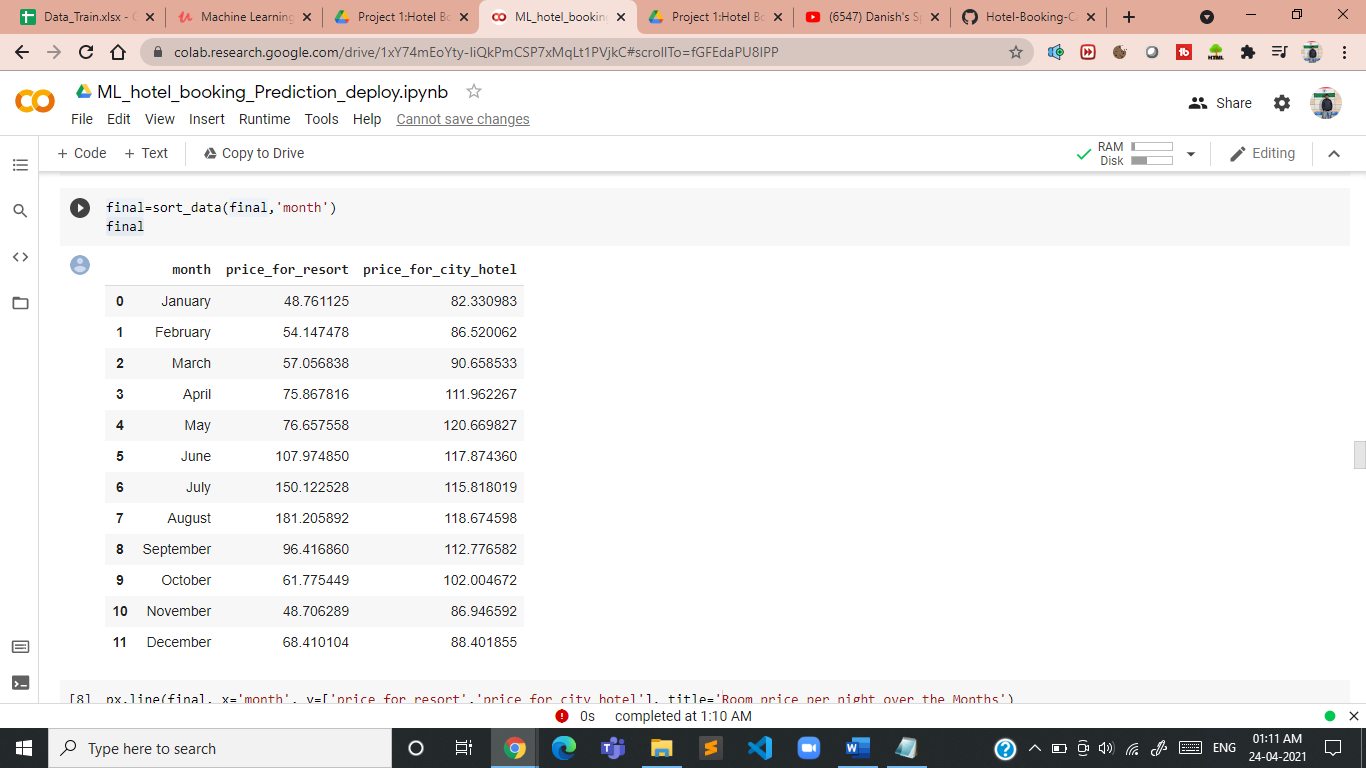
#### **Now we will observe over here is month column is not in order, & if we will visualise we will get improper conclusion**

#### **so very first we have to provide right hierarchy to the month column**

# !pip install sort-dataframeby-monthorweek

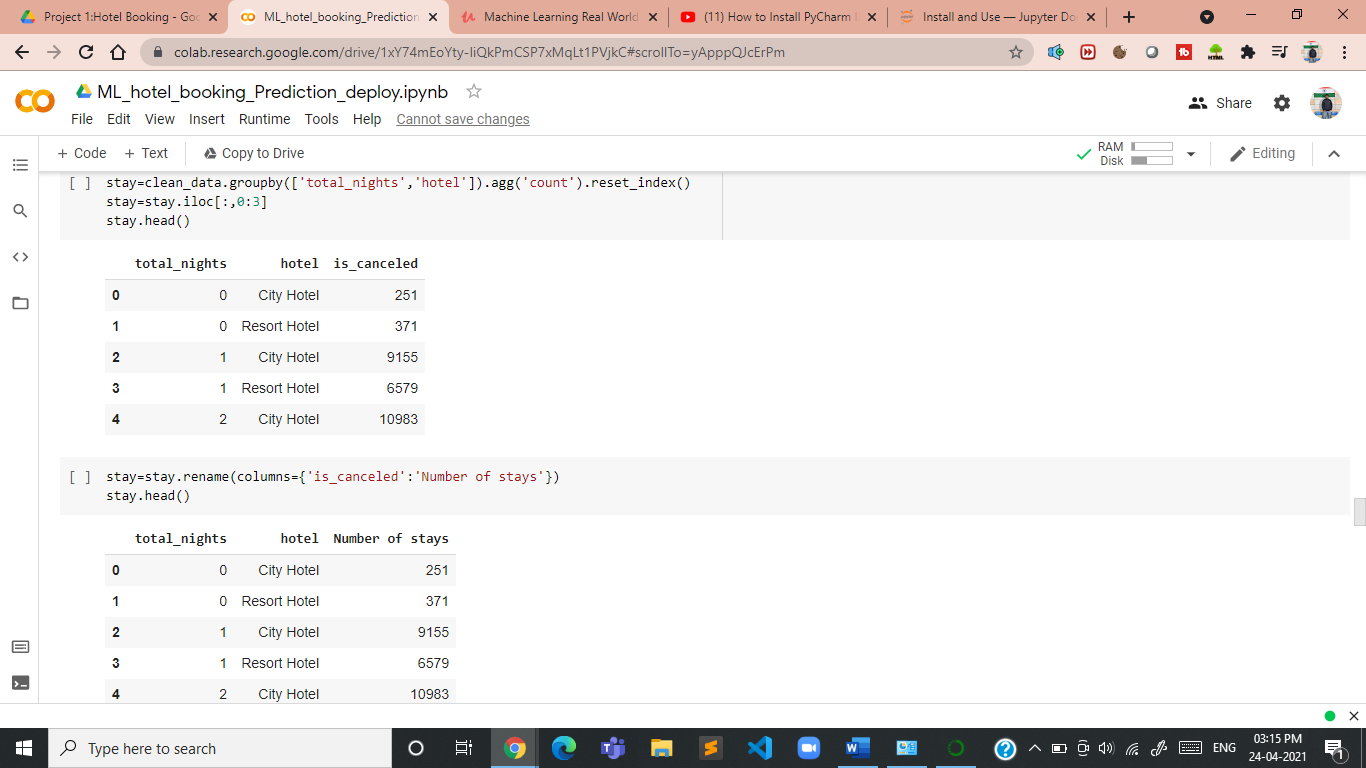
## Dependency package needs to be installed

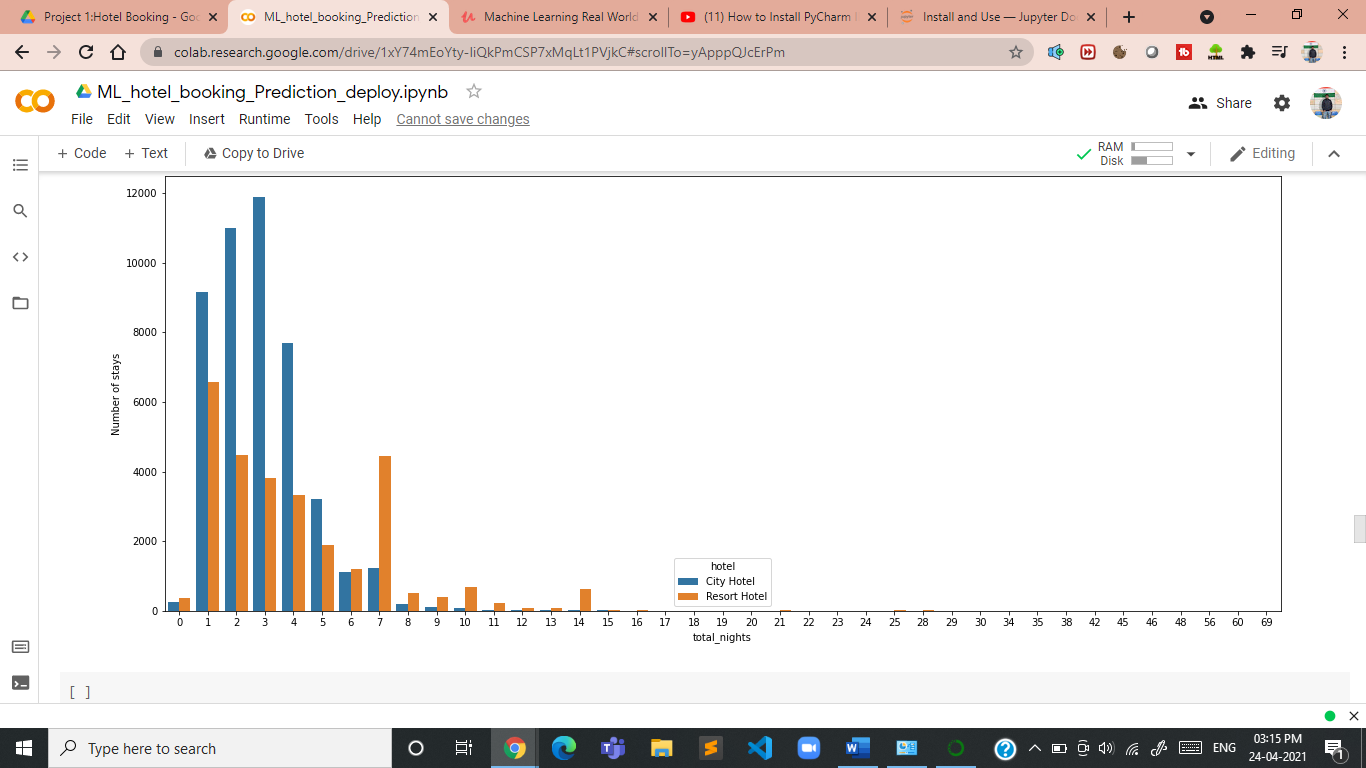
## pip install sorted-months-weekdays

****

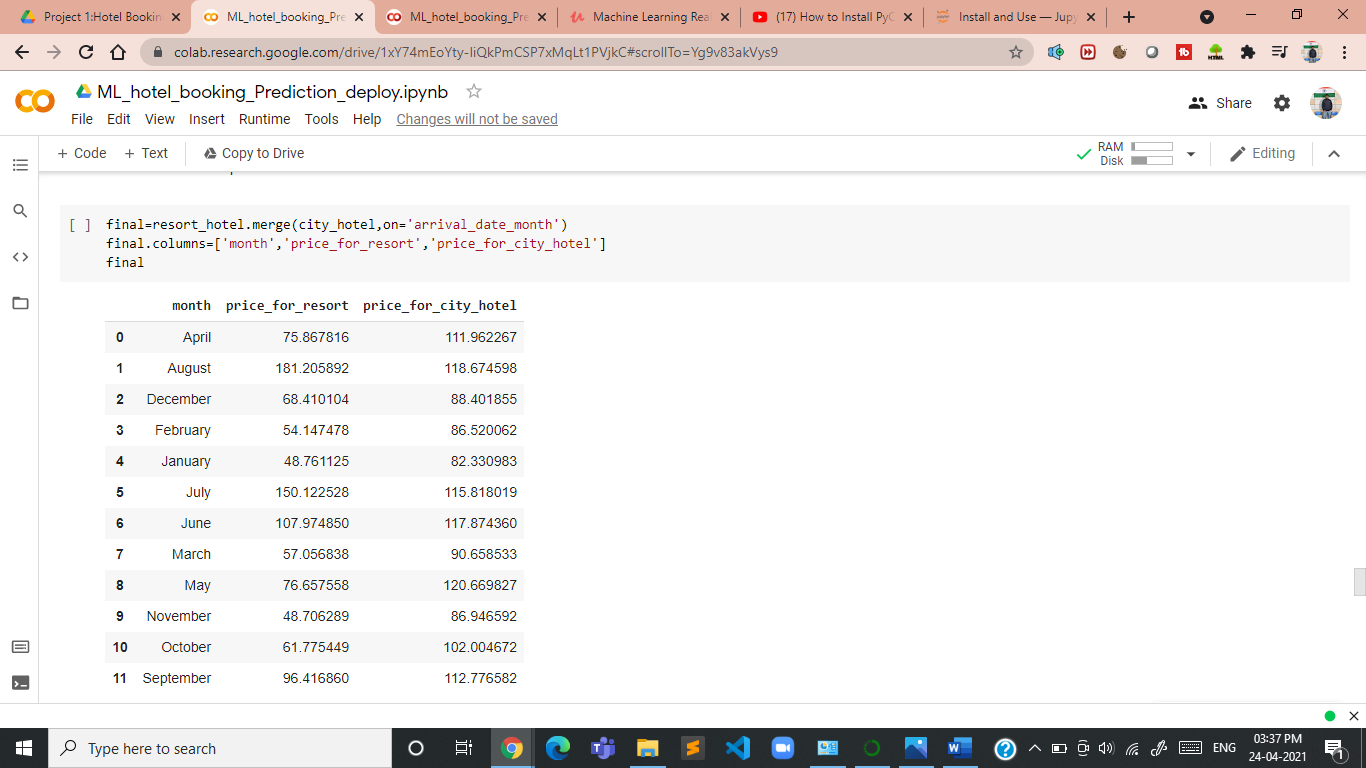
**Conclusion-->> This clearly shows that the prices in the Resort hotel are much higher during the summer (no surprise here)., The price of the city hotel varies less and is most expensive during spring and autumn.**

## How long do people stay at the hotels?



****

## How does the price per night vary over the year?

****

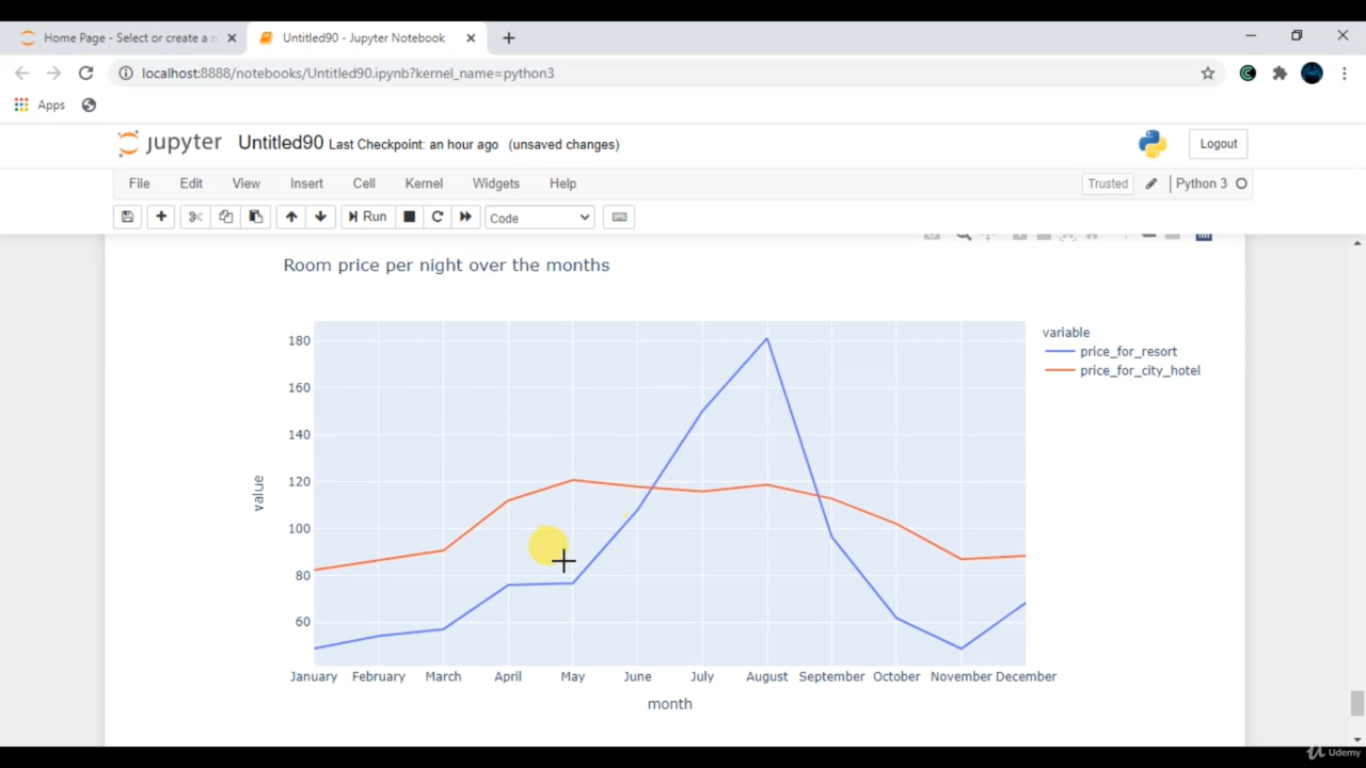
#### **now we will observe over here is month column is not in order, & if we will visualise we will get improper conclusion.**

#### **so very first we have to provide right hierarchy to the month column.**

**## !pip install sort-dataframeby-monthorweek**

**## Dependency package needs to be installed**

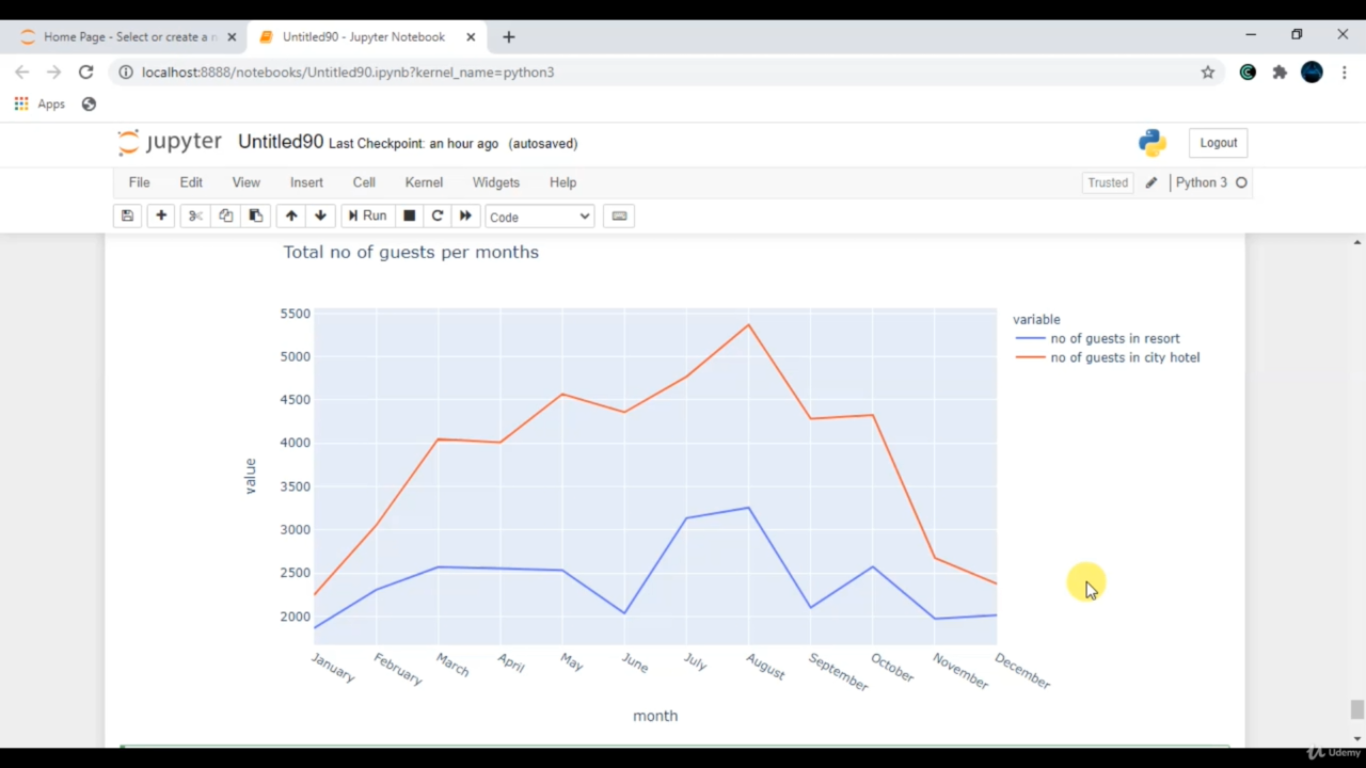
**## pip install sorted-months-weekdays**

****

### **Conclusion-->> This clearly shows that the prices in the Resort hotel are much higher during the summer (no surprise here)., The price of the city hotel varies less and is most expensive during spring and autumn.**

## Which are the most busy month or in which months Guests are high?

## 



### **Conclusion**

### The City hotel has more guests during spring and autumn, when the prices are also highest. In July and August there are less visitors, although prices are lower.   Guest numbers for the Resort hotel go down slighty from June to Septemer, which is also when the prices are highest. Both hotels have the fewest guests during the winter.

## How long do people stay at the hotels?

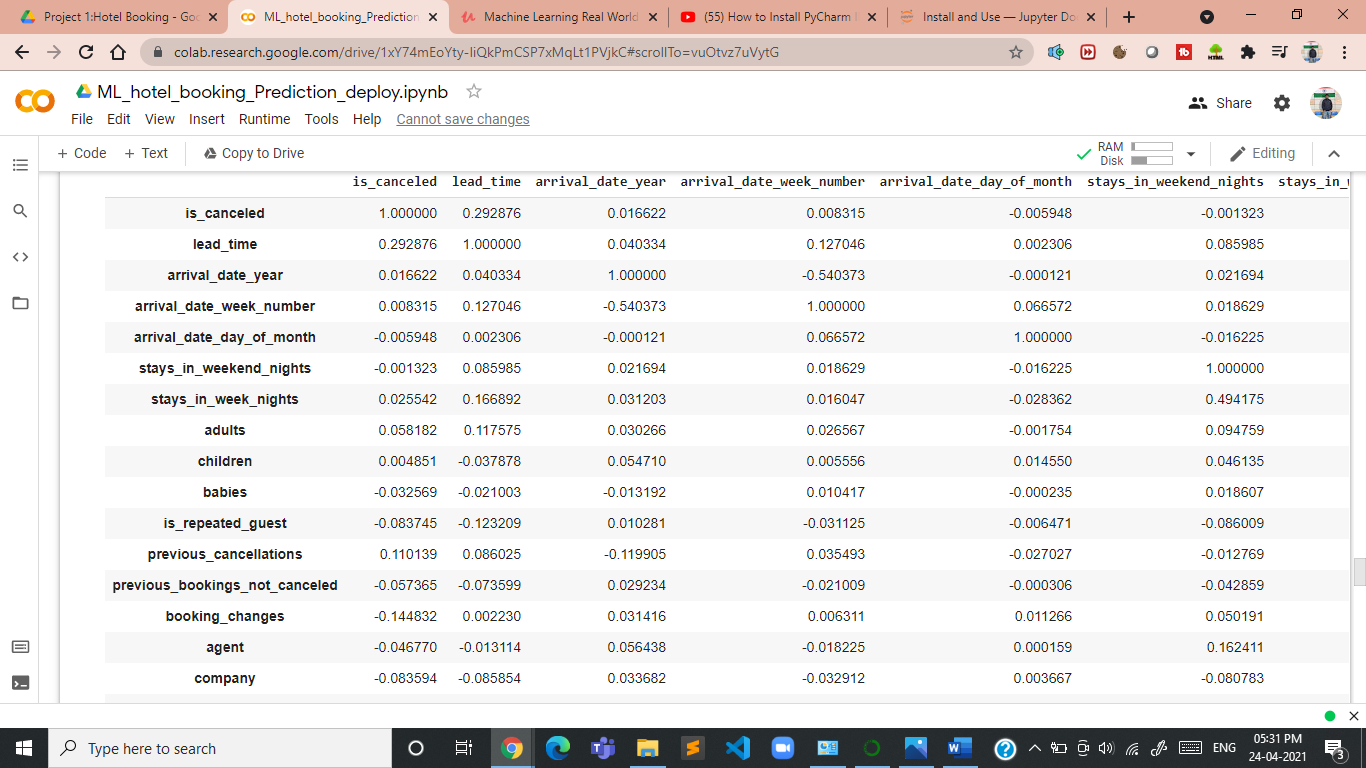
## Till now we have analyzed our data and we all understand our data in a very proper way. That what exactly the trend from where my other guests are, what exactly the distribution of the prize and how exactly the prices of a room that I am in, which month my guest is higher. We all have analyzed this trend as well.

## So that’s a time for your machine learning ascept.

## Now, the very first task with respect to your machine learning approach is exactly you have to select some important features using correlation concept for your machine learning model.

**co\_relation=data.corr()**

**co\_relation**

****

**co\_relation=data.corr()["is\_canceled"]**

**co\_relation**

**is\_canceled 1.000000**

**lead\_time 0.292876**

**arrival\_date\_year 0.016622**

**arrival\_date\_week\_number 0.008315**

**arrival\_date\_day\_of\_month -0.005948**

**stays\_in\_weekend\_nights -0.001323**

**stays\_in\_week\_nights 0.025542**

**adults 0.058182**

**children 0.004851**

**babies -0.032569**

**is\_repeated\_guest -0.083745**

**previous\_cancellations 0.110139**

**previous\_bookings\_not\_canceled -0.057365**

**booking\_changes -0.144832**

**agent -0.046770**

**company -0.083594**

**days\_in\_waiting\_list 0.054301**

**adr 0.046492**

**required\_car\_parking\_spaces -0.195701**

**total\_of\_special\_requests -0.234877**

## Name: is\_canceled, dtype: float64

After all the eng” these are all my categorical features that we have to take care of it.

**Index(['hotel', 'arrival\_date\_month', 'meal', 'market\_segment',**

**'distribution\_channel', 'reserved\_room\_type', 'deposit\_type',**

**'customer\_type', 'year', 'month', 'day', 'cancellation'],**

**dtype='object')**

**Now our aim is to derived some feature from the data.**

data\_cat['year']=data\_cat['reservation\_status\_date'].dt.year

data\_cat['month']=data\_cat['reservation\_status\_date'].dt.month

data\_cat['day']=data\_cat['reservation\_status\_date'].dt.day

**It means yoy have to convert these string data into some integer format using some feature encoding.**

**Now we are going to apply feature encoding technique.**

### **Perform Mean Encoding Technique**

cols=data\_cat.columns[0:8]

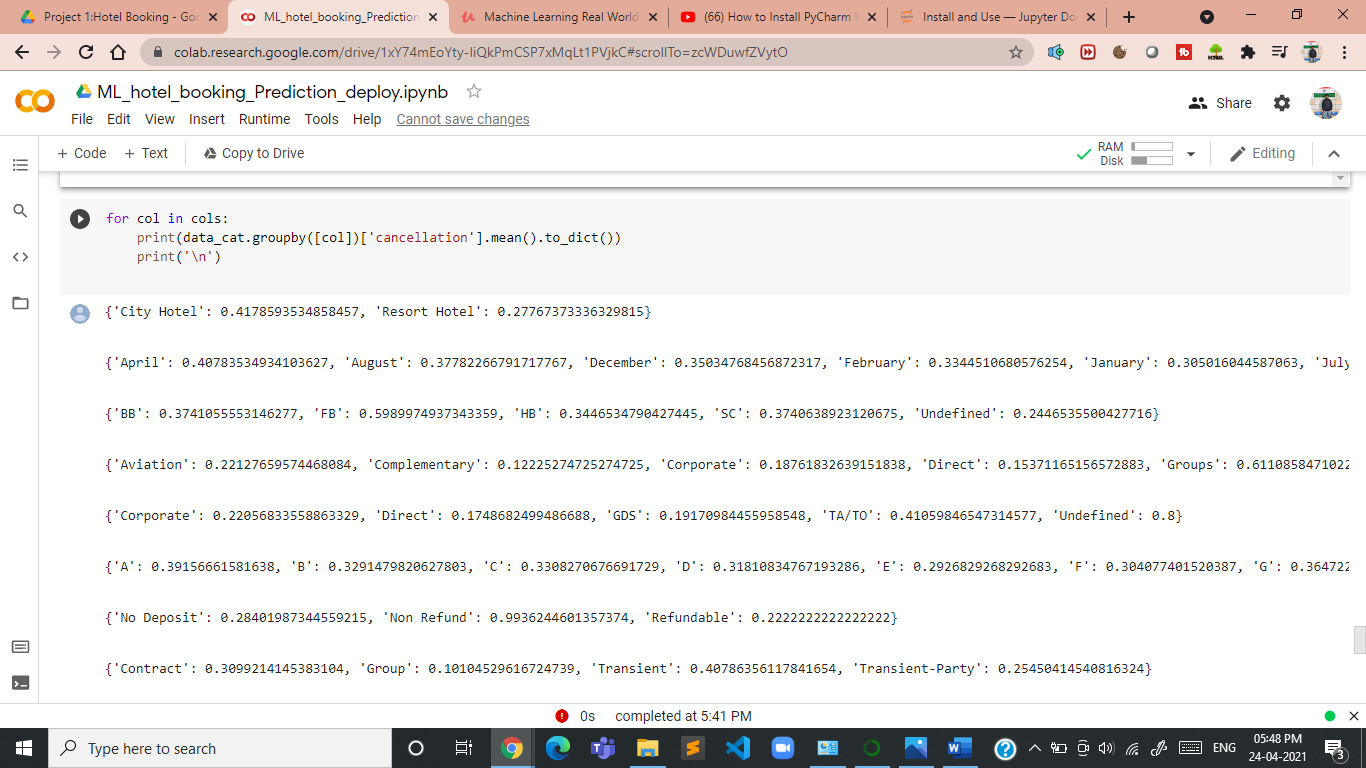
cols

Index(['hotel', 'arrival\_date\_month', 'meal', 'market\_segment',

'distribution\_channel', 'reserved\_room\_type', 'deposit\_type',

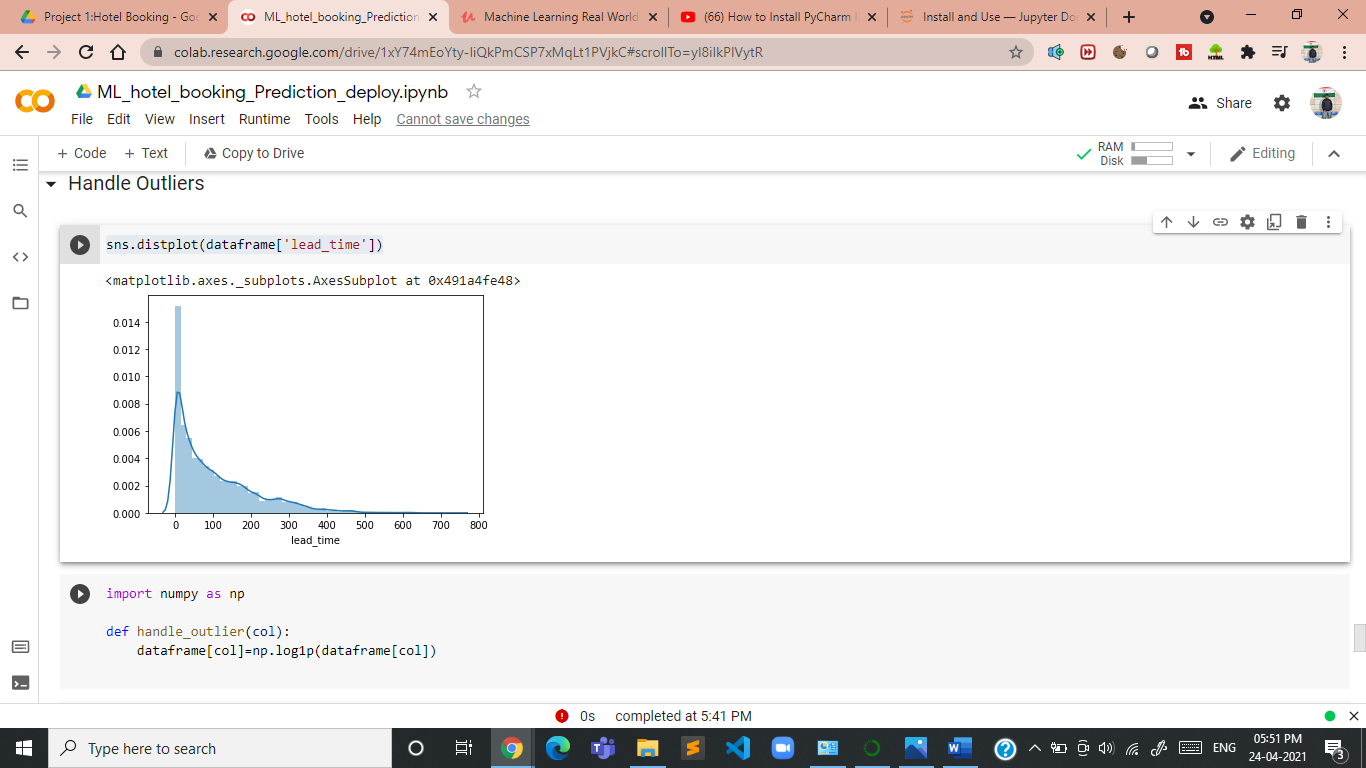
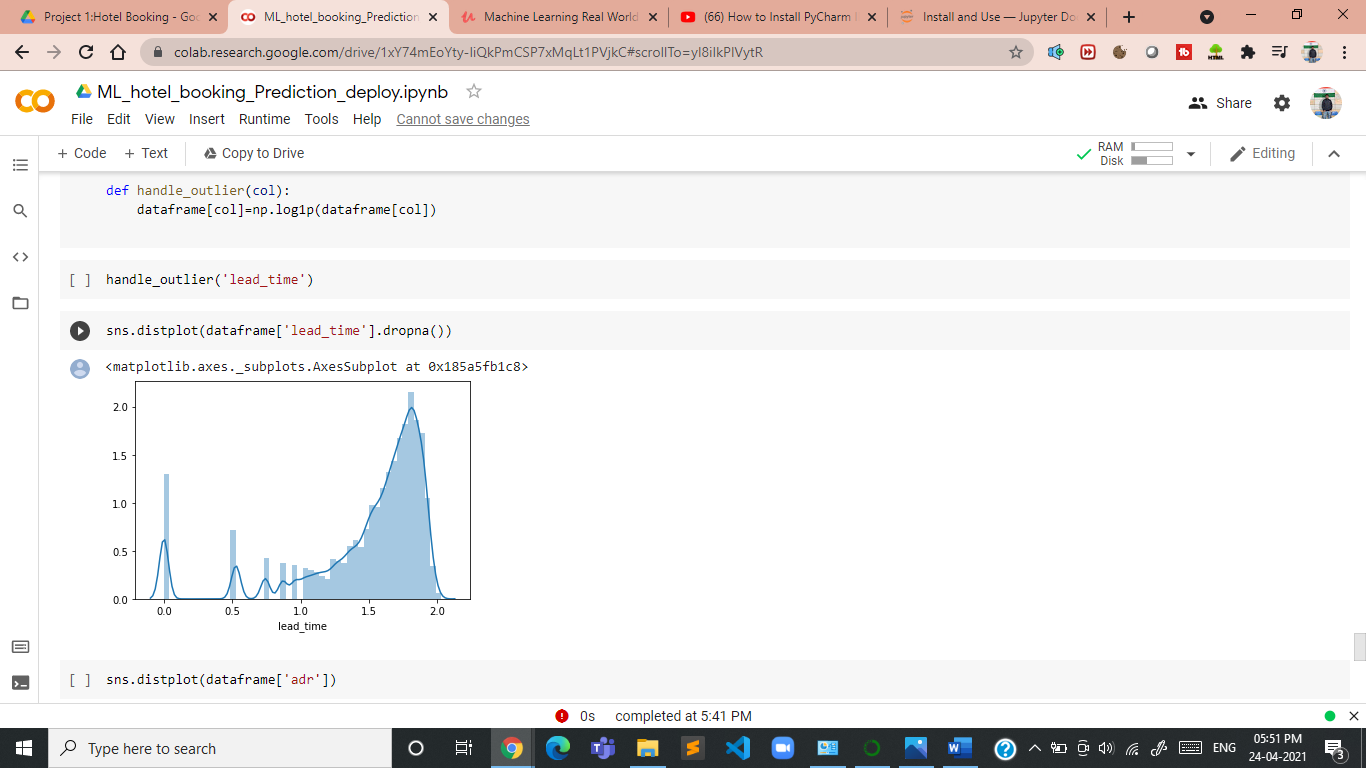
'customer\_type'],

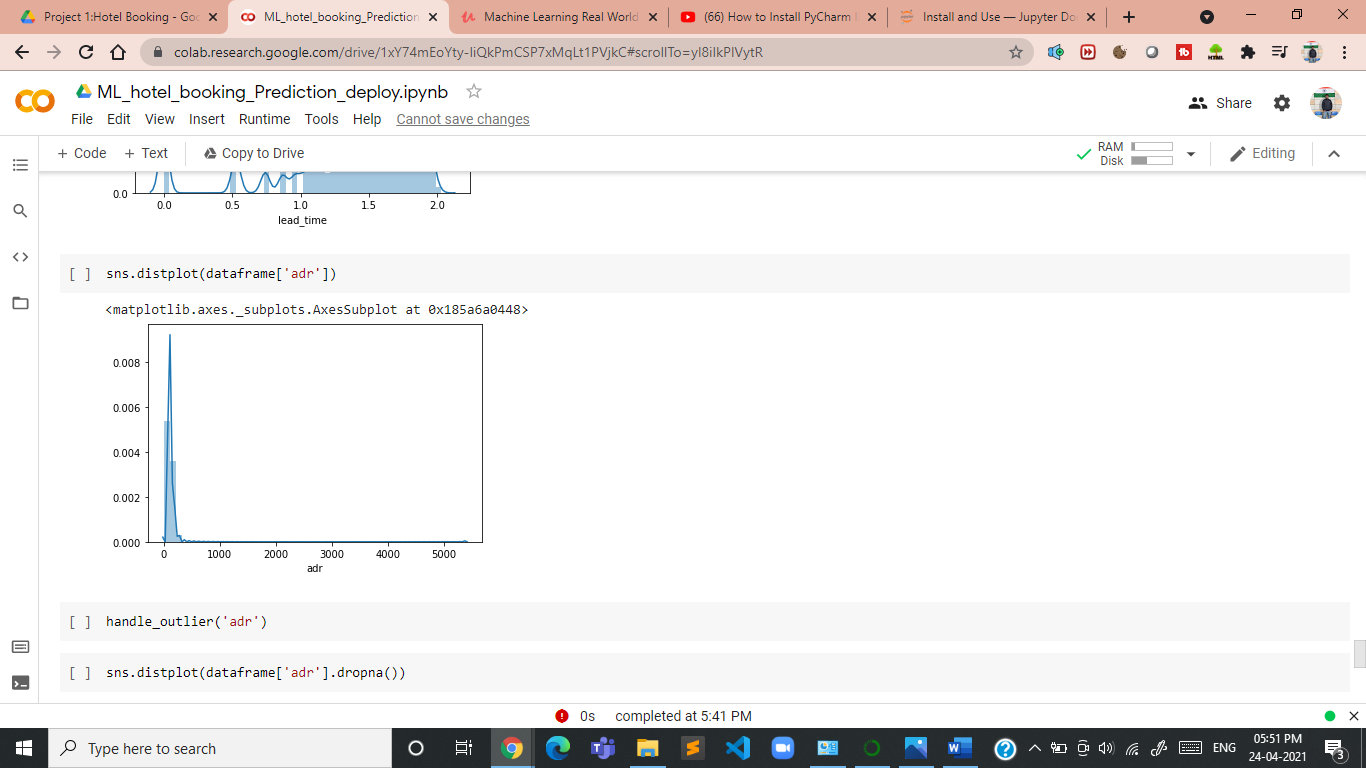
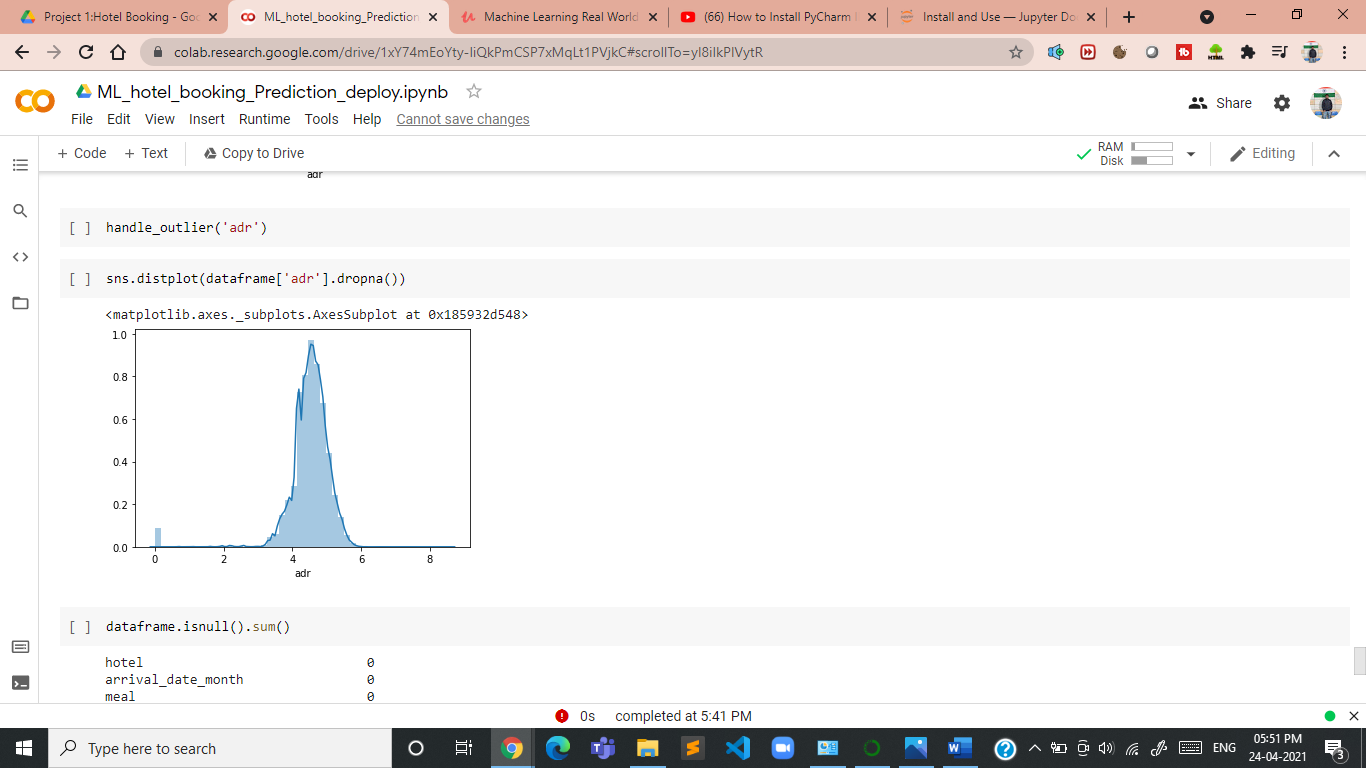
dtype='object')



**Now we Handle Outliers in our data**

**sns.distplot(dataframe['lead\_time'])**

** **

** **

*And this time you will see it is somehow very close to your normal distribution so now this is exactly that yoy really want for your machine learning purpose.*

**So, from this n number of feature you have to select some subset of features that are going to contribute more to my machine learning model.**

**# select a suitable alpha (equivalent of penalty).**

**# The bigger the alpha the less features that will be selected.**

**# remember to set the seed, the random state in this function.**

**# let's print the number of total and selected features**

**# this is how we can make a list of the selected features**

**selected\_feat = cols[(feature\_sel\_model.get\_support())]**

**print('total features: {}'.format((x.shape[1])))**

**print('selected features: {}'.format(len(selected\_feat)))**

**total features: 28**

**selected features: 16**

**selected\_feat**

**Index(['deposit\_type', 'year', 'month', 'day', 'lead\_time',**

**'arrival\_date\_week\_number', 'stays\_in\_week\_nights', 'adults',**

**'children', 'previous\_cancellations', 'previous\_bookings\_not\_canceled',**

**'booking\_changes', 'company', 'adr', 'required\_car\_parking\_spaces',**

**'total\_of\_special\_requests'],**

**dtype='object')**

Now it’s time to apply our machine learning algorithm on data beacause up to a greater extent, our data is ready to apply your machine learning algorithm on our data.

Now we apply machine learning algorithm.

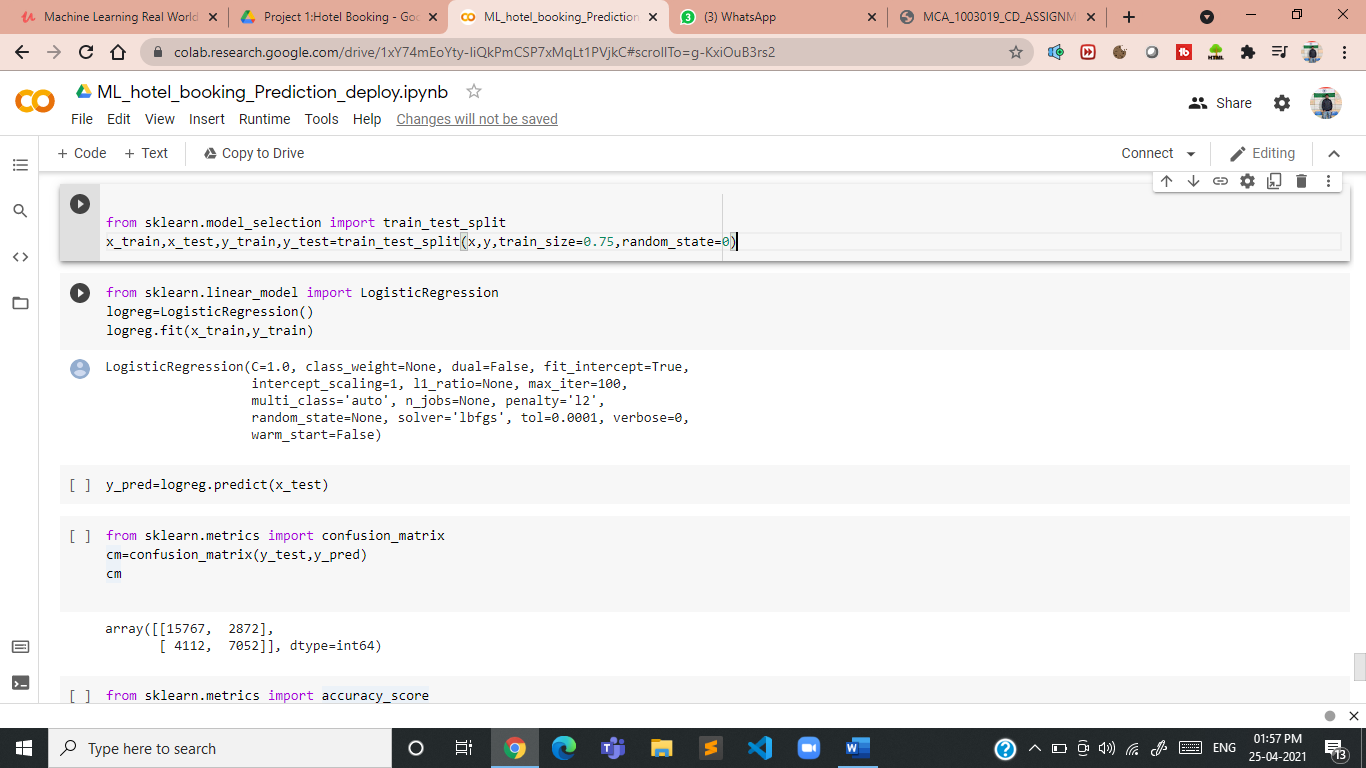
And then,

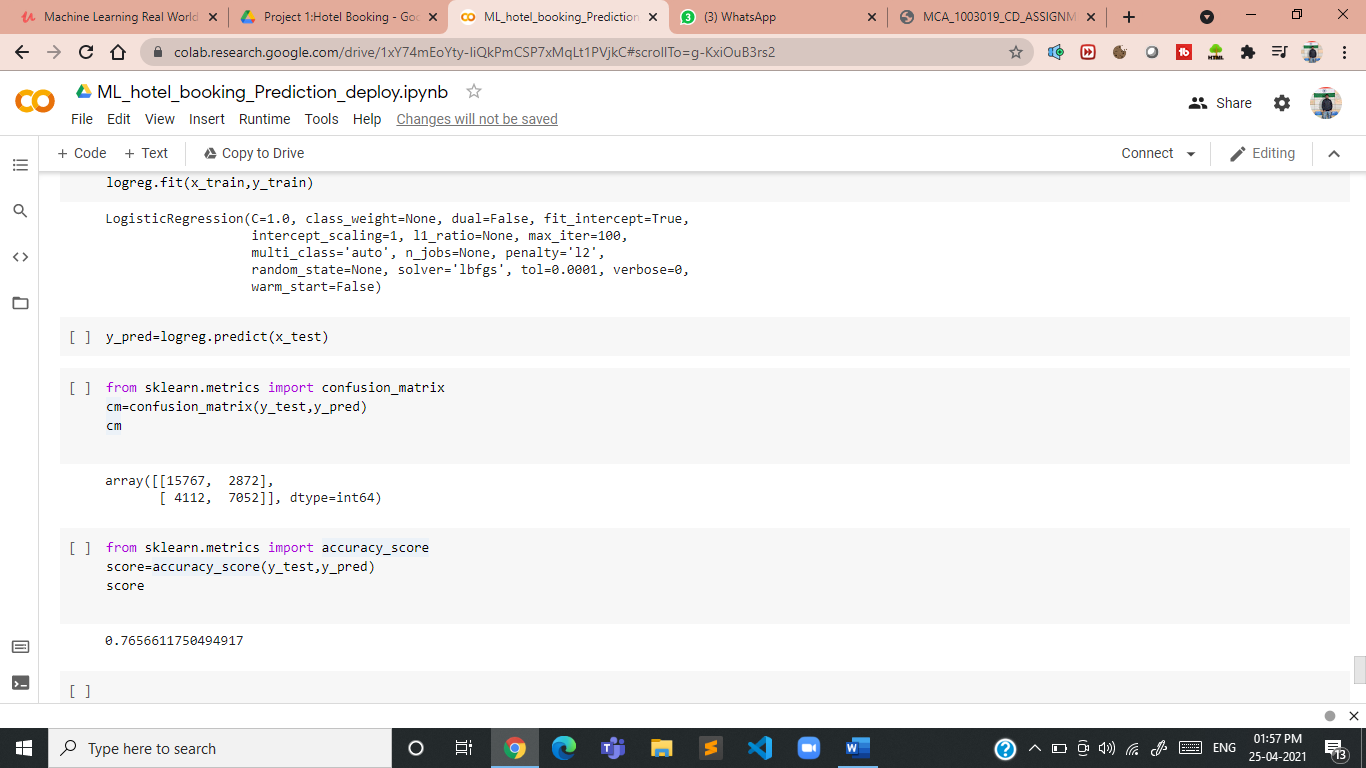
**Cross validate our model.**

Sor for that, what we have to do we very first split our data in the form of training as well as your testing, because once you have trained data, you can train your model.

And once you have test data and you can check what exactly is the accuracy of your model.

### **Splitting dataset & model Building**

****

****

**Now you will see my Logistic Regression model has somwhere close to seventy seven percent accuracy. But this is actually not my exact accuracy .**

**So how I can achieve my exact accuracy so in such scenarious you have to cross validate my model.**

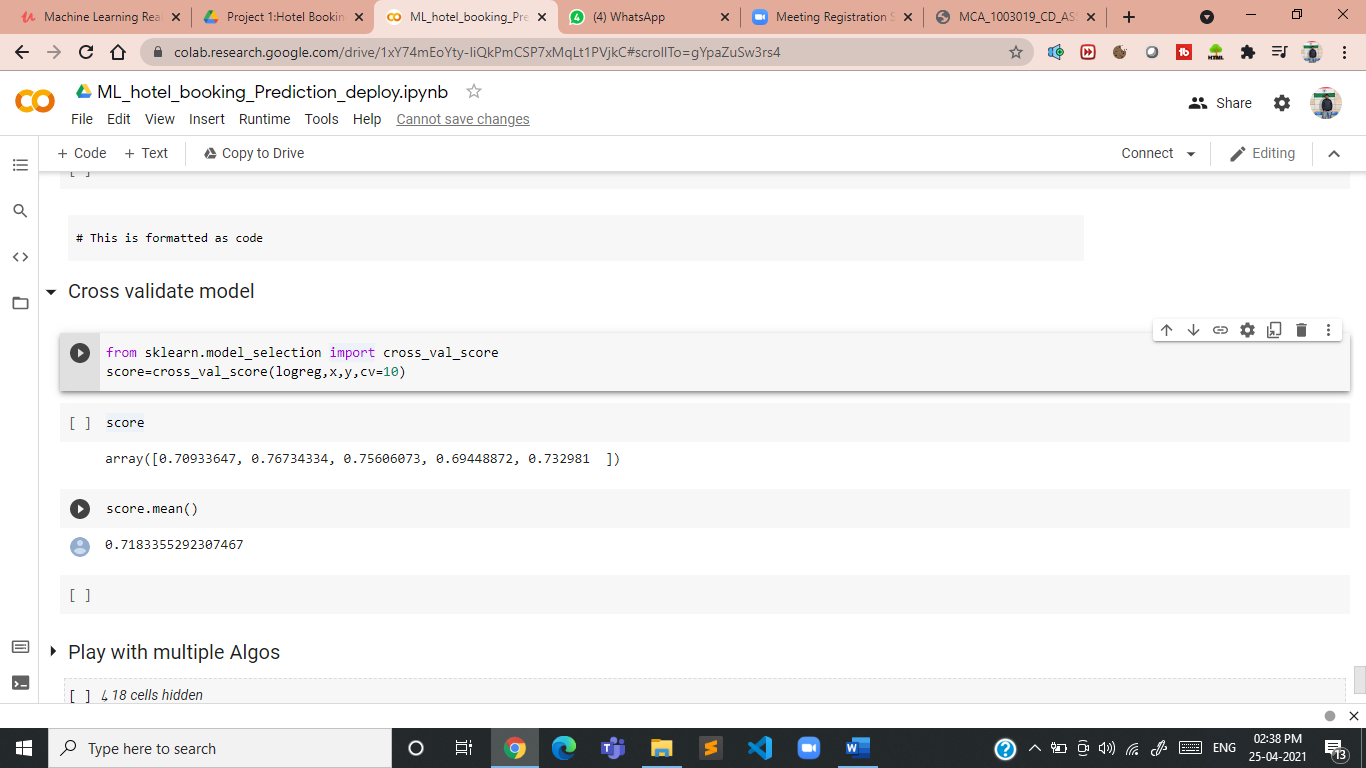
**If I change this random state value to any big number than the accuracy will be changed to any other percentage so than you will be thinking what is going here now in this we have to cross validate our model.**

**random\_state**

**So whenever you will get your accuracy without cross validate your model, you have to always say I have achieved accuracy within this,this range.**

**It means now to get the exact accuracy you have to cross, validate your model for this.**

### **Cross validate model**

****

**And now you can see over here it has accuracy of somewhere close to 73 percent. It means you’re 73 percent predictions are going to Correct.**

## How can this model be used?

### 1. This model will be helpful in highlighting the bookings which have high propensity of cancellation.

### 2. The hotels can use this model to impose high cancellation fee on bookings which the model has detected as likely to be cancelled. This will discourage the customer from cancelling or the booking will be done by a user who is less likely to cancel.

**I was able to predict the cancelation with 72.7% accuracy. I hope you find this project as interesting as I did!**

**Code-** [**https://da.gd/my\_code**](https://da.gd/my_code)