RAMAKRISHNA MISSION VIVEKANANDA EDUCATIONAL AND RESEARCH INSTITUTE

(Accredited by NAAC with 'A++' Grade)
Coimbatore – 641 020

in Collaboration with

Cognizant

SCHOOL OF MATHEMATICAL SCIENCE

SEPTEMBER – 2020

DEPARTMENT OF COMPUTER SCIENCE
CENTRE OF DATA SCIENCE

NAME	**: Swapnil Pradeep**
REG. NO	**: h19msds012**
PROGRAMME	**: M.Sc. (Data Science)**
SEMESTER	**: 2st Semester**
PROJECT TITLE	**: TAX EXTRATION**
COURSE CODE	**: 19DS2P06**

RAMAKRISHNA MISSION VIVEKANANDA EDUCATIONAL AND RESEARCH INSTITUTE

(Accredited by NAAC with 'A++' Grade)
Coimbatore – 641 020

in Collaboration with

Cognizant

SCHOOL OF MATHEMATICAL SCIENCE

SEPTEMBER – 2020

DEPARTMENT OF COMPUTER SCIENCE
CENTRE OF DATA SCIENCE

Bonafide Certificate

This is to certify that the project work done by _**Swapnil Pradeep**__ (**H19MSDS000**) entitled "**Project Title**" in his Second Semester during the academic year 2019-2020. Submitted for the Semester viva-voice Examination held on 14/09/2020 .

Staff In-Charge Head of the Department

Internal Examiner External Examiner

DECLARATION

I hereby declare that this project entitled "**PROJECT TITLE**" submitted to Ramakrishna Mission Vivekananda Educational and Research Institute, Coimbatore-20. is partial fulfillment of the requirement of the degree in M.Sc., (Data Science) is a record of the original project done by me under the guidance of Dr.R.Sridhar M.Sc., MCA., M.Phil., Ph.D., Professor & Head, Ramakrishna Mission Vivekananda Educational And Research Institute, Coimbatore - 641 020.

Place : Coimbatore

Signature of the Candidate

Date :14/09/2020

Swapnil Pradeep c

(H19MSDS012)

ACKNOWLEDGEMENT

My pranams to **Rev. Swami Garishthananda**, Secretary, Ramakrishna Mission Vidyalaya and **Rev. Swami Anapekshanada**, Asst. Administrative Head, FyCSAR, Ramakrishna Mission Vivekananda Educational and Research Institute, Coimbatore - 20 for providing me facilities and encouragement to complete the project work.

I express my profound gratitude to **Dr. R.Sridhar,** Professor & Head, FyCSAR, RKMVERI, Coimbatore-20 for his whole hearted encouragement and timely help.

I placed on record my heartfelt thanks and gratitude to **Sri. D. Raja,** Project Consultant and **Sri. V.Dineshkumar,** Assistant Professor, FyCSAR, RKMVERI, Coimbatore-20 under whose guidance the work has been done. It is a matter of joy that without his valuable suggestions, able guidance, scholarly touch and piercing insight that he offered me in each and every stage of this study, coupled with his unreserved, sympathetic and encouraging attitude, this thesis could not have been presented in this manner.

I am thankful to the following members from the industry for their multi-dimensional review and support to complete my project work.

1. Rohini Krishnan, Senior Manager, Cognizant
2. Manjunaath Alagiriswamy, Senior Associate, Cognizant
3. Muralidharan Sivakumar, Senior Associate, Cognizant
4. Kulshum Azmi, Behavioural Trainer, Cognizant
5. Rekha Priyadharshini, Python SME, Cognizant
6. Rajasekar, Pyhton SME, Cognizant
7. Hemnath Muthukrushnun, Data Scientist SME, Cognizant
8. Shabarivasan RC, Data Scientist SME, Cognizant
9. Balasubramanian Mahadevan, Managing Partner, Algolitics India LLP, Coimbatore

SYNOPSIS

This is a text processing project which consist of customer data of paying hotel rooms, and here I want to extract the price which the customer has paid the tax for the hotel rooms. There are three types of tax paid by customer which we can see in data set, also there is another column which consist onsite rate which mean full price include all tax. The process is to extract the tax price and calculate with the onsite price and then find the tax exclude price which will give information that how much tax has been paid by the customer.

TAX EXTRACTION

STRATEGIC GOAL

*to know how much tax is to be paid by customer

OBJECTIVE

*Recommend all types of tax information for customer

IMPORTING LIBRARIES

In [2]:

```
import re
import pandas as pd
import spacy
```

In [3]:

```
pd.set_option('display.max_colwidth', 300)
```

IMPORTING DATA SET

In [4]:

```
taxdata = pd.read_csv('IDeaS - Tax Type Analysis.csv')
```

In [5]:

```
taxdata.head()
```

Out[5]:

	refid	hotelcode	websitecode	dtcollected	ratedate	los	guests	roomtype	onsiterate
0	7309709	350853	5	07-12-2019 19:16	28-12-2019	1	1	Suite-Two Bedrooms Sea View with Jetted Tub (Maria)	584.70
1	7136380	111808	5	07-12-2019 19:16	24-12-2019	1	1	King Suite	375.45
2	7309709	350853	5	07-12-2019 19:16	28-12-2019	1	1	Suite-Two Bedrooms Sea View with Jetted Tub (Maria)	660.30
3	8026928	73457	5	07-12-2019 19:16	15-01-2020	1	1	Double comfort	164.58
4	1768789	492908	2	07-12-2019 19:16	23-12-2019	1	1	Two-Bedroom House	88.00

5 rows × 28 columns

TAKING TAX TYPE 1

In [6]:

```
taxdata1 = taxdata[taxdata['taxstatus']==1]
```

In [7]:

```
taxdata1.head()
```

Out[7]:

	refid	hotelcode	websitecode	dtcollected	ratedate	los	guests	roomtype	onsiterat
0	7309709	350853	5	07-12-2019 19:16	28-12-2019	1	1	Suite-Two Bedrooms Sea View with Jetted Tub (Maria)	584.
2	7309709	350853	5	07-12-2019 19:16	28-12-2019	1	1	Suite-Two Bedrooms Sea View with Jetted Tub (Maria)	660.
4	1768789	492908	2	07-12-2019 19:16	23-12-2019	1	1	Two-Bedroom House	88.
9	9021247	170339	5	07-12-2019 19:16	10-02-2020	1	1	Double or Twin Room	52.
13	6732632	638539	5	07-12-2019 19:16	13-12-2019	1	1	One-Bedroom Apartment with Terrace	158.

5 rows × 28 columns

SPLITING THE TAX TYPE 1 VALUE

spliting each numerical value from this text for further calulation of tax exclude amount

In [8]:

```
txt = 'City tax € 2.57, Government Tax (Pay at the property) € 1.50, VAT € 66.80'
```

In [9]:

```
tmp = txt.split(',')
```

In [10]:

```
len(tmp)
```

Out[10]:

```
3
```

In [11]:

```
def SplitTaxText(txt):
    ls = []
    for tx in txt.split(','):
        ls1 = tx.split('€')
        if len(ls1) == 1:
            ls1 = tx.split('CHF')
            if len(ls1) == 1:
                ls1 = tx.split('DKK')
                ls.append(ls1)
            else:
                ls.append(ls1)
        else:
            ls.append(ls1)
    return ls
```

spliting all values in the column of tax type

In [12]:

```
taxdata1['taxsplited'] = [SplitTaxText(txt) for txt in taxdata1['taxtype']]
```

```
C:\Users\swapy\anaconda3\lib\site-packages\ipykernel_launcher.py:1: Settin
gWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

In [13]:

```
taxdata1.head()
```

Out[13]:

	refid	hotelcode	websitecode	dtcollected	ratedate	los	guests	roomtype	onsiterat
0	7309709	350853	5	07-12-2019 19:16	28-12-2019	1	1	Suite-Two Bedrooms Sea View with Jetted Tub (Maria)	584.
2	7309709	350853	5	07-12-2019 19:16	28-12-2019	1	1	Suite-Two Bedrooms Sea View with Jetted Tub (Maria)	660.
4	1768789	492908	2	07-12-2019 19:16	23-12-2019	1	1	Two-Bedroom House	88.
9	9021247	170339	5	07-12-2019 19:16	10-02-2020	1	1	Double or Twin Room	52.
13	6732632	638539	5	07-12-2019 19:16	13-12-2019	1	1	One-Bedroom Apartment with Terrace	158.

5 rows × 29 columns

In [14]:

```
taxdata1.to_csv('TaxStatus1.csv')
```

calulating the tax exclude rate in a single row

In [15]:

```
def CalculateTax(onsiteRate, taxText):
    print(onsiteRate)
    print(taxText)
    resRate = onsiteRate
    taxrate = 0
    for ls in taxText:
        if ls[0].find('(Pay at the property)') < 0:
            print(ls)
            if len(ls) > 0:
                if str(ls[1]).find('%') < 0:
                    resRate = resRate - float(ls[1])
                else:
                    taxrate = str(ls[1]).replace('%', '')
                    resRate = resRate - resRate * float(taxrate)/100
    return resRate
```

In [16]:

```
ls = [['City tax' , '2.57%'], [ 'Government Tax (Pay at the property)' , 1.50], [ 'VAT'
, 66.80]]
rate = 584.7
```

In [17]:

```
CalculateTax(rate, ls)
```

```
584.7
[['City tax', '2.57%'], ['Government Tax (Pay at the property)', 1.5], ['V
AT', 66.8]]
['City tax', '2.57%']
['VAT', 66.8]
```

Out[17]:

```
502.87321000000003
```

In [18]:

```
# Load Tax Status 1 data
taxdata1 = pd.read_csv('TaxStatus1.csv')
```

calculating rate for every row

In [19]:

```
def CalculateTaxforEveryRow(row):
    # updating the value of the row
    print(row)
    row['AmountExcludeTax'] = CalculateTax(row['onsiterate'], row['taxsplited'])
    return row
```

In [20]:

```
taxdata1[['onsiterate', 'taxsplited']]
```

Out[20]:

	onsiterate	taxsplited
0	584.7	[['City tax ', ' 2.57'], [' Government Tax (Pay at the property) ', ' 1.50'], [' VAT ', ' 66.80']]
1	660.3	[['City tax ', ' 2.90'], [' Government Tax (Pay at the property) ', ' 1.50'], [' VAT ', ' 75.46']]
2	88.0	[['"Included in bungalow price:9 % VAT"], ["BGN 0.98 City tax per person per night'"], [' includes taxes and charges']]
3	52.0	[['City tax (Pay at the property) ', ' 4.00'], [' VAT ', ' 4.36']]
4	158.5	[['City tax (Pay at the property) ', ' 0.50'], [' VAT ', ' 14.36']]
...	...	...
72336	117.4	[['City tax (Pay at the property) ', ' 2.40'], [' VAT ', ' 10.45']]
72337	57.0	[['"Included:7 % VAT"], ["& 1.30 City tax per person per night'"], [' includes taxes and charges']]
72338	139.0	[['City tax (Pay at the property) ', ' 3.00'], [' VAT ', ' 12.36']]
72339	97.5	[['VAT ', ' 8.30'], [' Cleaning Fee (Pay at the property) ', ' 20.00']]
72340	60.0	[['"Included:10 % VAT"], ['15 % Property service charge'], ["Breakfast'"], [' includes taxes and charges']]

72341 rows × 2 columns

In [21]:

```
# Call CalculateTax for every row
```

In [ ]:

```
taxdata1['CalculatedTax'] = taxdata1['onsiterate'] - taxdata1['AmountExcludeTax']
```

In [26]:

```
taxdata1.to_csv('TaxStatus1_final.csv')
```

In [ ]:

```

```