IBM Applied Data Science Capstone Report

Opening a Take-Away Indian Restaurant in Christchurch City of New Zealand

DECEMBER 19, 2020

Authored by: Pranav Shekhawat

"The problem isn't finding data, it's figuring out what to do with it" –Mike Loukides

Table of Content

1. Introduction

I am always fascinated by food and particularly Indian food. You can travel anywhere in the world you will find at-least one Indian restaurant. While the taste and quality might differ from place to place and restaurant to restaurant it has never stopped me from trying every Indian restaurant in town.

Christchurch is one of the major cities in New Zealand which is in redevelopment phase since the major earthquake in 2011. It is rapidly growing and attracts many student and immigrants to come and settle down in the city. Indian population in Christchurch has steadily grown in the past few years.

These past couple of months as the world is suffering from Coronavirus it has changed the way people live their lives. It has affected many businesses especially restaurants around the world. They had to change the way they operate which in most cases have increased the cost of running the restaurants. Some restaurants might not be able to keep up with cost of additional cleaning and maintaining high standards for customers and in turn shut down. People nowadays think twice before going out to eat in public because of the virus and may end up ordering food online using UberEats.

At the moment there are around 20 Indian Restaurants in the city most of which are franchise of the same. While the cost of running a take-away restaurants will be much lower compared to a dine-in restaurant, moreover there are no Indian Restaurants which solely focus on Take-Away, for those reasons my suggestion is to open a Take-Away Indian restaurant. Future expansion will be much easier as well if one decides to turn it into a franchise. (*The business details will be decided by the investor or entrepreneur*).

2. Business Problem/Objective

Question: Which locations would you recommended in the city of Christchurch if someone is looking to open or invest in a take-away Indian restaurant?

This project will solely focus on analyzing and selecting best locations in the city of Christchurch to open a Take-away Indian restaurant. We will be using data science and machine learning techniques.

Target Audience: The project findings will be useful for Investors, Entrepreneurs, Food Enthusiast Researchers looking to get into the restaurant business

(Please Note: My suggestion to open a take-away restaurant is solely based on the current changes around the world due to coronavirus. And people ordering online rather than dine-in. The investors can use this research to open a dine-in restaurants as well**)**.

3. Data Description

Data Requirement:

The requirement for this project are:

- List of Suburbs for the city of Christchurch in New Zealand.
- Geographical location such as longitude and latitude data of Christchurch and all the suburbs
- List of Venue categories in the city
- Venue Data relating to Indian restaurants such as Location

Data source and Use:

With the help of **Beautiful-soup** package in python a web-scrapping technique, we will obtain data for "List of Suburbs name" from the Wikipedia page:
(https://en.wikipedia.org/wiki/Category:Suburbs_of_Christchurch).
Further cleaning of this data was done on the local network to work as required for the project.
Further with the help of Geocoder package we will retrieve geographical coordinates i.e. Longitude and Latitude for each suburb in the city.
The data about different venues in different Suburbs of the city will be obtained by using **Foursquare** location data. Foursquare has one of the largest database of locations with over 100 million venues. The data includes venue names, locations, menus, photos, review and more. As such, the Foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.
The data acquired from Foursquare API will have the information of venues within a specified radius of the longitude and latitude of the suburbs. The data will be as follows:

- **Suburbs**: Name of the Suburbs
- **Suburbs Latitude**: Latitude of the Suburbs
- **Suburbs Longitude**: Longitude of the Suburbs
- **Venue**: Name of the Venue
- **Venue Latitude**: Latitude of Venue
- **Venue Longitude**: Longitude of Venue
- **Venue Category**: Category of Venue

The data described above will have substantial information to build our model. Later we will be clustering the suburbs together based on our objective. We then present our observations and findings. With the help of those findings our stakeholders can take the necessary decision.

4. Methodology

First, we will import and install necessary libraries and packages to get started with our project, additional libraries will be installed if needed for use through the process:

- **Pandas**: To collect and manipulate data in JSON and HTML and then data analysis
- **requests**: Handle http requests
- **Beautiful-Soup**: library to parse HTML and XML documents
- **matplotlib**: Detailing the generated maps
- **folium**: Generating maps of London and Paris
- **sklearn**: To import **Kmeans** which is the machine learning model that we are using.

We will explore each suburb in the city, plot the map to show the suburbs being considered and then build our model by clustering and at a later stage plot the new map with the clusters of Indian restaurants in city suburbs. This way we can visualize and examine the findings.

Second, In the data extraction stage, we begin with collecting the required data for the suburbs of Christchurch City. We need data that includes the name of each suburb in the city for now. To extract the data for Christchurch, we scrape the Wikipedia page to take the list of suburbs, the data had a lot of mistake so it had to be extracted to a local hard drive and processed properly to work as required. The data set has 71 records each representing a suburb:

```
Data.head()
```

	Suburbs
0	Addington, New Zealand
1	Aidanfield
2	Aranui
3	Avondale, Christchurch
4	Avonhead

Third, we needed to extract geographical coordinates for the city as well as for each of the suburbs. We used geocoder package and defined a function to retrieve the data. We looped the function to perform the task for each suburb.

Function:

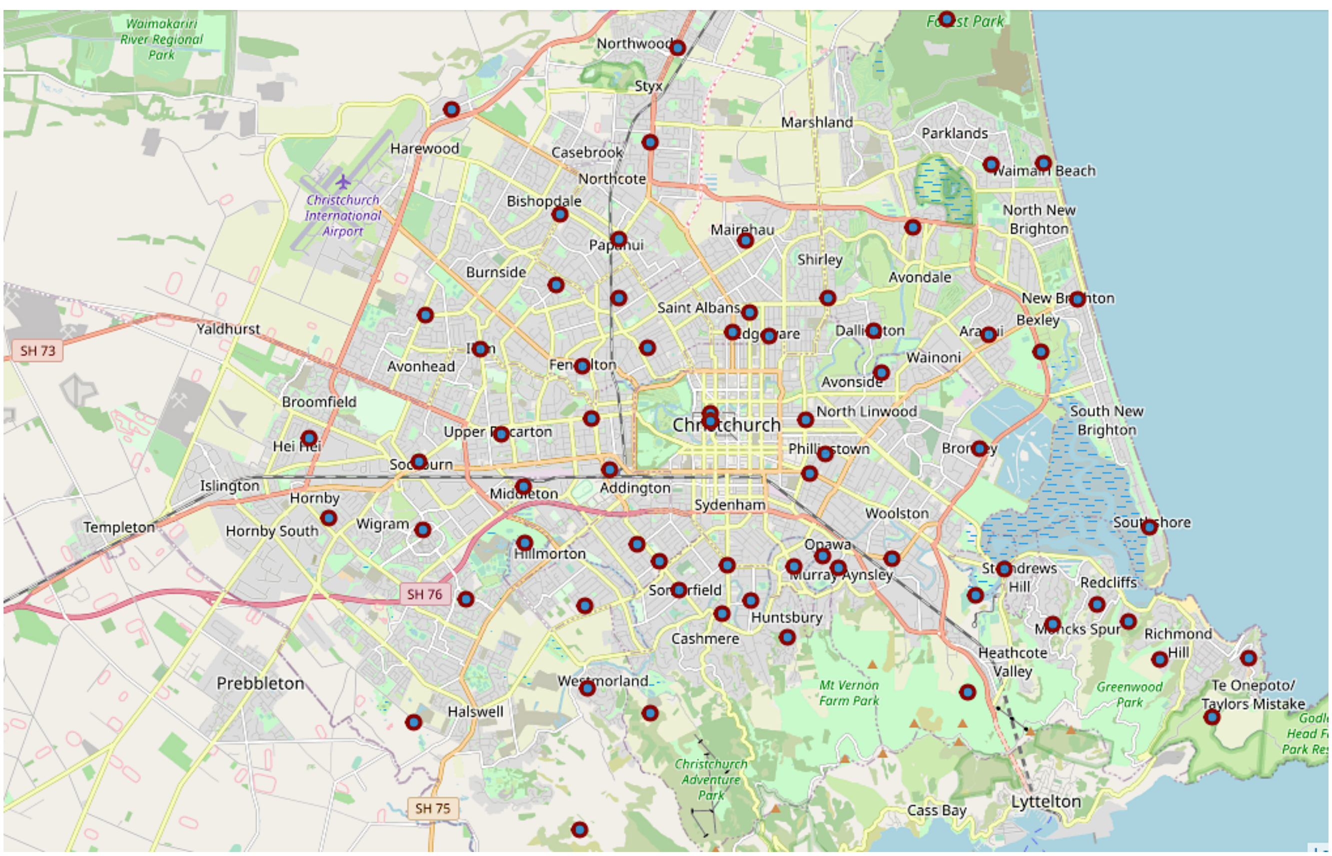
```python
def get_latilong(suburbs):
    lati_long_coords = None
    while(lati_long_coords is None):
        g = geocoder.arcgis('{}, Christchurch, New Zealand'.format(suburbs))
        lati_long_coords = g.latlng
    return lati_long_coords

get_latilong('Christchurch Central, Christchurch')
```

Output:

```python
Locate # Checking if we have retrived coordinates
```

```
[[-43.53964999999994, 172.60590000000002],
 [-43.564340277999975, 172.56806022000012],
 [-43.51393977899994, 172.7054613350001],
 [-43.530347448999976, 172.6324268940001],
 [-43.51032999999995, 172.55741000000012],
 [-43.521204498999964, 172.67728250100004],
```

After we have combined the location data set with our suburb data set we move on to plotting and visualizing a **map of Christchurch City** with suburb data using matplotlib library.



Fourth, after we have visualized and confirmed the suburbs and their location on the map we move on to find out what each Suburb is like and what are the common venue and venue categories within 800-meter radius.
This is where we use Foursquare API. With the help of Foursquare we define a function which collects information pertaining to each suburb including Suburb name, Coordinates (Longitude and Latitude), Venue name, Venue categories and Venue Coordinates (Longitude and Latitude).

Foursquare API

```
radius = 800
LIMIT = 100

venues = []

for lat, long, suburbs in zip(Data['Latitude'], Data['Longitude'], Data['Suburbs']):

    # create the API request URL
    url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}".format(
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        long,
        radius,
        LIMIT)

    # make the GET request
    results = requests.get(url).json()["response"]['groups'][0]['items']

    # return only relevant information for each nearby venue
    for venue in results:
        venues.append((
            suburbs,
            lat,
            long,
            venue['venue']['name'],
            venue['venue']['location']['lat'],
            venue['venue']['location']['lng'],
            venue['venue']['categories'][0]['name']))
```

Output

(1330, 7)

	Suburbs	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Addington, New Zealand	-43.53965	172.6059	Tower Junction Mega Centre	-43.538805	172.606176	Shopping Mall
1	Addington, New Zealand	-43.53965	172.6059	The Court Theatre	-43.541363	172.610309	Theater
2	Addington, New Zealand	-43.53965	172.6059	Addington Coffee Co-op	-43.543590	172.611595	Coffee Shop
3	Addington, New Zealand	-43.53965	172.6059	North & South Gourmet (南北小厨)	-43.543909	172.611512	Asian Restaurant
4	Addington, New Zealand	-43.53965	172.6059	Christchurch Train Station	-43.539805	172.608015	Train Station

```
Venue_data.tail()
```

	Suburbs	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
1325	Woolston, New Zealand	-43.5566	172.6801	The Tannery - Boutique Retail & Arts Emporium	-43.557187	172.680316	Boutique
1326	Woolston, New Zealand	-43.5566	172.6801	Blue Smoke	-43.556940	172.680281	Wine Bar
1327	Woolston, New Zealand	-43.5566	172.6801	In Situ	-43.558668	172.673332	Café
1328	Woolston, New Zealand	-43.5566	172.6801	Three Boys Brewery	-43.555746	172.678998	Brewery
1329	Woolston, New Zealand	-43.5566	172.6801	Mitchelli's	-43.557833	172.680027	Italian Restaurant

As we are looking for Indian restaurants in the city, first we check for total number of unique categories of venue and check if the category contains "**Indian restaurant**". Secondly, if it does then we check the total number of Indian restaurants.

Unique Venue Categories

```
print('There are {} uniques categories.'.format(len(Venue_data['VenueCategory'].unique())))
```

```
There are 152 uniques categories.
```

```
Venue_data['VenueCategory'].unique()[:150]
```

```
array(['Shopping Mall', 'Theater', 'Coffee Shop', 'Asian Restaurant',
       'Train Station', 'Stadium', 'Falafel Restaurant',
       'Afghan Restaurant', 'Liquor Store', 'Bar', 'Pet Store',
```

Total Number of Indian Restaurant

```
"Indian Restaurant" in Venue_data['VenueCategory'].unique()
```

```
True
```

```
a=pd.Series(Venue_data.VenueCategory) # Checking top 10 venue categories
a.value_counts()[:20]
```

```
Café                   166
Hotel                   77
Bar                     68
Park                    59
Coffee Shop             45
Restaurant              37
Thai Restaurant         34
Italian Restaurant      28
Gastropub               27
Supermarket             26
Shopping Mall           23
Indian Restaurant       22
```

Now we move on **to One Hot Encoding** method to investigate our categorical datatype of the venue categories. This helps to convert the categorical data into numeric data.
We won't be using label encoding in this situation since label encoding might cause our machine learning model to have a bias or a sort of ranking which we are trying to avoid by using One Hot Encoding.

One Hot Encoding

```
# one hot encoding
Chch_onehot = pd.get_dummies(Venue_data[['VenueCategory']], prefix="", prefix_sep="")

# add suburbs column back to dataframe
Chch_onehot['Suburbs'] = Venue_data['Suburbs']

# move suburbs column to the first column
fixed_columns = [Chch_onehot.columns[-1]] + list(Chch_onehot.columns[:-1])
Chch_onehot = Chch_onehot[fixed_columns]
Chch_grouped = Chch_onehot.groupby('Suburbs').mean().reset_index()
Chch_onehot.head(5)
```

Output

	Suburbs	Accessories Store	Afghan Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop	BBQ Joint
0	Addington, New Zealand	0	0	0	0	0	0	0	0	0
1	Addington, New Zealand	0	0	0	0	0	0	0	0	0
2	Addington, New Zealand	0	0	0	0	0	0	0	0	0

After preforming One Hot Encoding we then calculate the **mean of the frequency of occurrence** of each venue category for each suburb.

```
Chch_grouped = Chch_onehot.groupby('Suburbs').mean().reset_index()
print(Chch_grouped.shape)
Chch_grouped
```

(72, 153)

	Suburbs	Accessories Store	Afghan Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Resta
0	Addington, New Zealand	0.000000	0.030303	0.000000	0.000000	0.000000	0.09
1	Aidanfield	0.000000	0.000000	0.000000	0.000000	0.000000	0.0C
2	Aranui	0.000000	0.000000	0.000000	0.000000	0.000000	0.0C
3	Avondale, Christchurch	0.000000	0.000000	0.000000	0.010989	0.010989	0.0'
4	Avonhead	0.000000	0.000000	0.000000	0.000000	0.000000	0.0C
5	Avonside	0.000000	0.000000	0.000000	0.000000	0.000000	0.0C

At **last** we move on to building a model using our machine learning techniques Kmeans Clustering from sklearn library. We will be generating 3 clusters using just the data frame consisting of suburbs and mean frequency occurrence of Indian restaurants. We will allocate every data point to its nearest centroid while keeping centroid as small as possible. We will add labels (0 to 2) to every data point based on centroid. This result will allow us to make our observation and conclusions.

Data Frame (Shape: 71, 2)

	Suburbs	Indian Restaurant
0	Addington, New Zealand	0.000000
1	Aidanfield	0.000000
2	Aranui	0.000000
3	Avondale, Christchurch	0.010989
4	Avonhead	0.000000

Kmeans Clustering

```
# set number of clusters
kclusters = 3

Chch_clustering = Chch_IR.drop(["Suburbs"], 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Chch_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)
```

Output

	Suburbs	Indian Restaurant	Labels
0	Addington, New Zealand	0.000000	1
1	Aidanfield	0.000000	1
2	Aranui	0.000000	1
3	Avondale, Christchurch	0.010989	1
4	Avonhead	0.000000	1

We then visualize the map and examine our clusters by expanding on our code using the Cluster Labels column, which we will discuss in the result and discussion section.

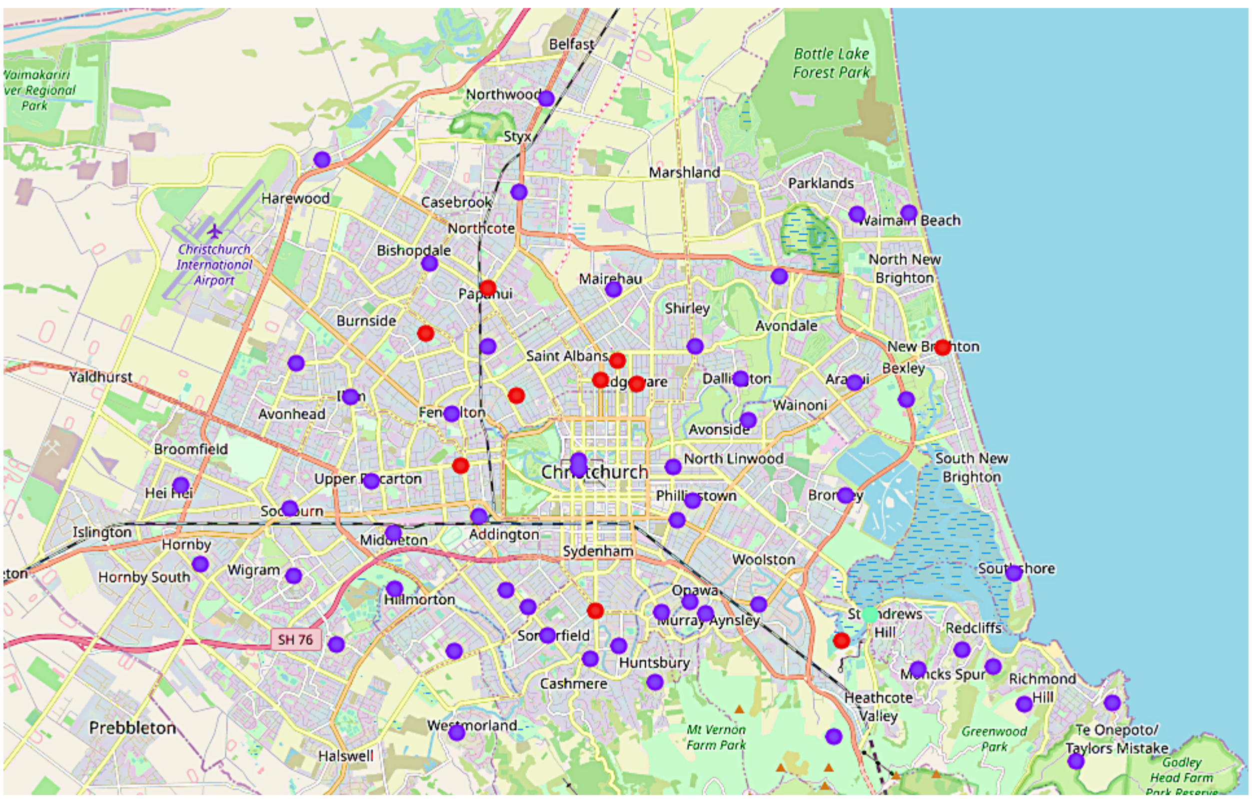
5. Result

The results obtained from our machine learning method K-means clustering, Christchurch suburbs were divided into 3 cluster based on the frequency of occurrence of Indian restaurants.

After Examining the clusters, we can visualize that (for high to low occurrence):

- Cluster 0 (Red) contains highest number of Indian Restaurants
- Cluster 2 (Green) contains moderate number of Indian Restaurants
- Cluster 1 (Purple) Contains very few or no Indian Restaurants

Map of Christchurch City with Cluster and Labels



6. Discussion

The cluster map points out that **Cluster 0** and **Cluster 2** have most of the Indian restaurant, and they do not cover the central part of the city which is currently in the redevelopment phase. On the other hand, we can see that **Cluster 1** has few to no Indian restaurants. This cluster can be called the cluster of opportunity. This cover about **~80%** of the suburbs in the city of Christchurch which include the central part of the city. For personal experience most of the students local and international live in the city center as well has working professional. Investors can invest but not recommended in **Cluster 0** and **Cluster 2** if they have enough capital power and are ready to take intense competition.

Therefore, I can recommend to investors, entrepreneurs or anyone interested who are looking to open a take-away Indian restaurant to capitalize on these findings. As mentioned before **Clusters 1** covers the central part of the city which has plenty of opportunity as it is under redevelopment phase and new spaces as well as land is up for sale/rent.

The Suburbs name as per cluster for reference are as follows:

Cluster 0

```
Chch_merged.loc[Chch_merged['Labels'] == 0]
```

	Suburbs	Indian Restaurant	Labels	Latitude	Longitude
37	Merivale	0.050000	0	-43.516545	172.615832
21	Edgeware	0.062500	0	-43.514293	172.647681
49	Riccarton, New Zealand	0.037037	0	-43.529800	172.601190
23	Ferrymead	0.090909	0	-43.563577	172.702118
13	Bryndwr	0.076923	0	-43.504451	172.591626
59	St Albans, New Zealand	0.100000	0	-43.513430	172.638040
44	Papanui	0.035714	0	-43.495825	172.608419
51	Richmond, Christchurch	0.047619	0	-43.509690	172.642450
42	New Brighton, New Zealand	0.090909	0	-43.507330	172.728670
64	Sydenham, New Zealand	0.052632	0	-43.557870	172.636630

Cluster 1

```
Chch_merged.loc[Chch_merged['Labels'] == 1]
```

	Suburbs	Indian Restaurant	Labels	Latitude	Longitude
45	Parklands, New Zealand	0.000000	1	-43.481640	172.706000
50	Richmond Hill, New Zealand	0.000000	1	-43.575697	172.750407
43	Opawa	0.000000	1	-43.555987	172.661955
47	Redcliffs	0.000000	1	-43.565306	172.733929
48	Redwood, Christchurch	0.000000	1	-43.477430	172.616600
41	Murray Aynsley Hill	0.000000	1	-43.558363	172.665880
40	Mount Pleasant, New Zealand	0.000000	1	-43.568966	172.722116
39	Moncks Bay	0.000000	1	-43.568532	172.742179
46	Phillipstown, New Zealand	0.000000	1	-43.536644	172.662645
0	Addington, New Zealand	0.000000	1	-43.539650	172.605900
38	Middleton, New Zealand	0.000000	1	-43.542944	172.583200
54	Sockburn, New Zealand	0.000000	1	-43.538000	172.555660
55	Somerfield, New Zealand	0.000000	1	-43.562480	172.624003

Cluster 2

```
Chch_merged.loc[Chch_merged['Labels'] == 2]
```

	Suburbs	Indian Restaurant	Labels	Latitude	Longitude
60	St Andrews Hill	0.2	2	-43.55853	172.70962

7. Limitation

This project was limited to frequency of occurrence Indian Restaurants in Christchurch suburbs. There are a lot of factors that could be considered for further studies and projects such as, population per suburb, popularity of a particular cuisine in the city, landmarks near-by which may attract more people in a particular suburb. A lot of these data is not readily available for use or might require some field work in a bigger scale. Use of Premium API calls and more advanced software will help in deeper analysis.

8. Conclusion

The Project achieved what it set out to achieve by exploring the suburbs of Christchurch city and with the help of clustering we could point out the best locations for new take-away Indian restaurant. While there are certain limitations to the project but this will help in further studies in a positive way. Researcher can pick a particular suburb and dive into much more details.

To Conclude my recommendation to investors, entrepreneurs and food enthusiast that Cluster 1 is that best way forward for them, and they and can capitalizes these findings as per their needs.

(Please Note: My suggestion to open a take-away restaurant is solely based on the current changes around the world due to coronavirus. And people ordering online rather than dine-in. The investors can use this research to open a dine-in restaurants as well.**).**

END.