LUCKY COUNTRY (AUSTRALIA), CLUSTERING SIMILAR CITIES BY VENUE CATEGORIES

1. Australia is a big country with small population.

2. Australia is an international tourist attraction.

3. Travelers a~~rou~~nd the world know few of Australian cities

4. Travelers need to know what to expect from a new city.

5. Grouping of cities give a valuable answer to "What to expect from new city"

ter Australian_cities_comparison Last Checkpoint: 2 hours ago (autosaved)

lit View Insert Cell Kernel Widgets Help

Run Code

[20]:
```
df_city_raw.iloc[0:5] #Cluster 1 of similar cities
```

[20]:

	Significant Urban Areas	State	Population2019	Population2011	Growth	percent_of_national_population
0	Sydney	New South Wales	4914343	4231954	+19.10%	20.93%
1	Melbourne	Victoria	4893870	3999982	+24.08%	19.86%
2	Brisbane	Queensland	2430180	2065996	+19.20%	9.85%
3	Perth	Western Australia	2045479	1728867	+19.12%	8.24%
4	Adelaide	South Australia	1340794	1262940	+6.56%	5.38%

[21]:
```
df_city_raw.iloc[96:101] #Cluster 2 of similar cities
```

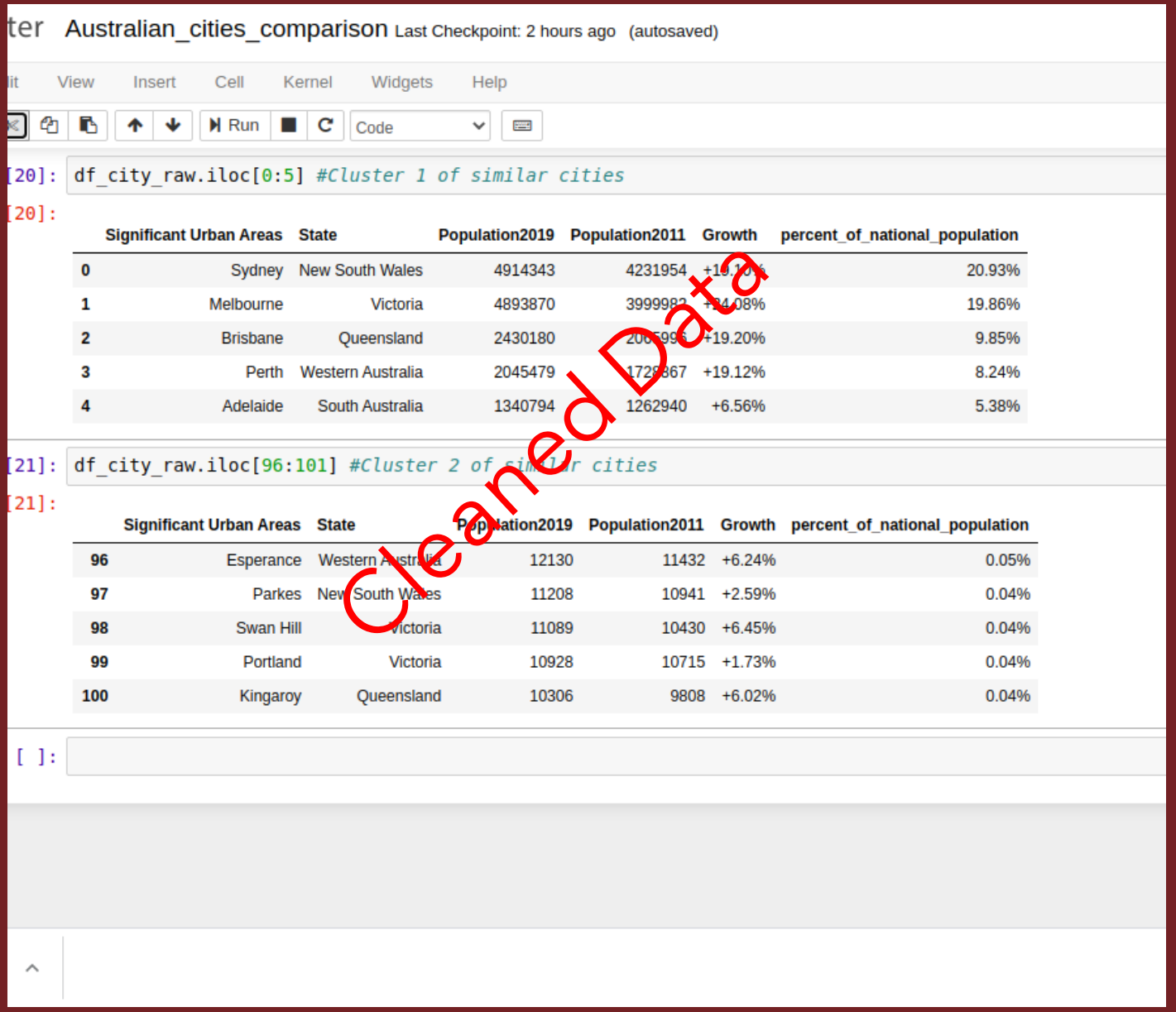
[21]:

	Significant Urban Areas	State	Population2019	Population2011	Growth	percent_of_national_population
96	Esperance	Western Australia	12130	11432	+6.24%	0.05%
97	Parkes	New South Wales	11208	10941	+2.59%	0.04%
98	Swan Hill	Victoria	11089	10430	+6.45%	0.04%
99	Portland	Victoria	10928	10715	+1.73%	0.04%
100	Kingaroy	Queensland	10306	9808	+6.02%	0.04%

[ ]:

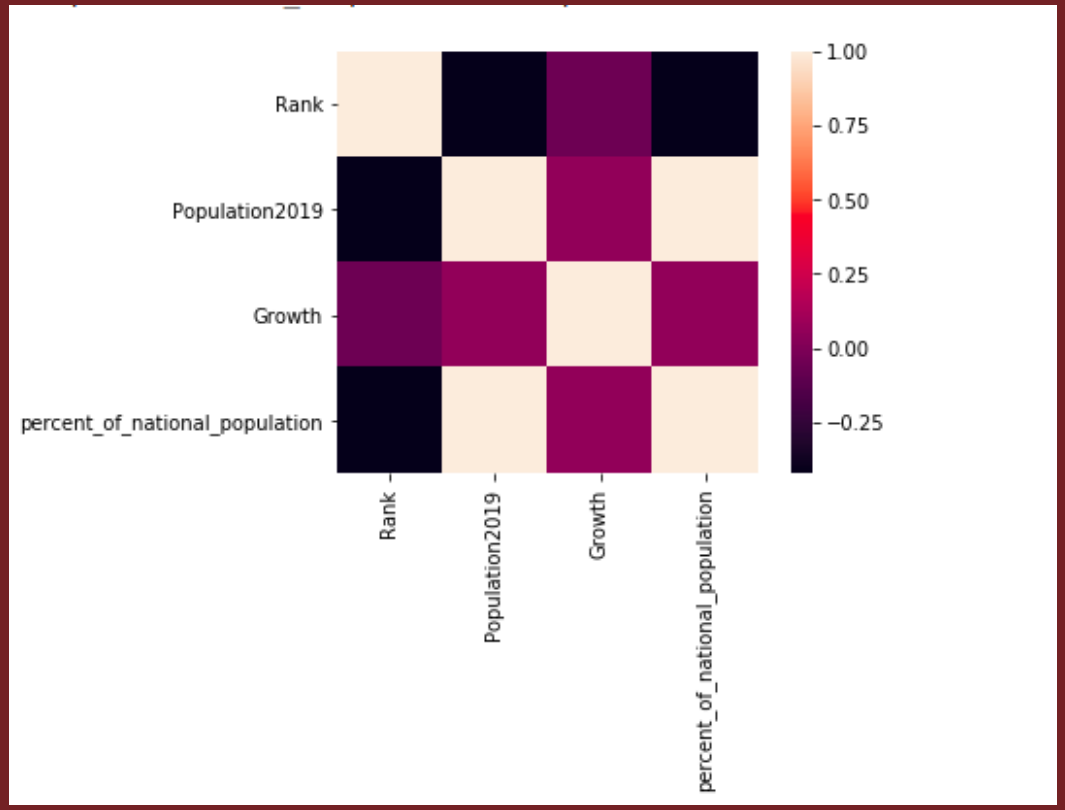
Data acquisition:

1. Get the list of Australian cities from Wikipedia
2. Get the city latitude and longitude from Geopy API
3. Get the venue list for cities from Foursquare API



Data cleaning:

1. Clean Wikipedia data and remove unwanted symbols from the data
2. Convert numbers in string format to float/numeric values.
3. Apply one hot encoding for categorical values.



Exploratory Data Analysis:

1. Correlation analysis clearly shows which columns can be dropped
2. percent_of_national_population can be removed as it does not add any new information.
3. There are total of 477 venue categories and one state column.

Thus a total of 478 categorical features which are reduced to numeric values using one hot encoding.

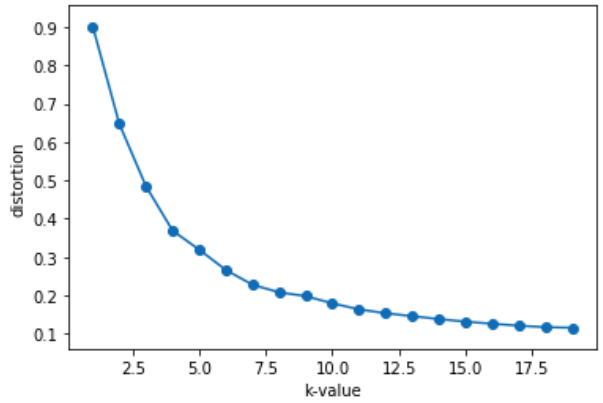
jupyter Australian_cities_comparison Last Checkpoint: an hour ago (autosaved)

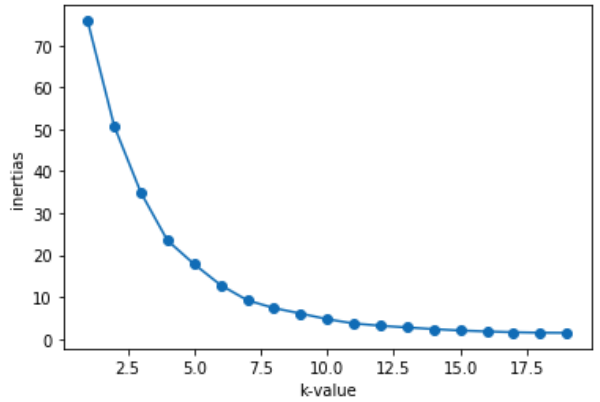
File Edit View Insert Cell Kernel Widgets Help

Run Code

```
plt.show()

df_best_k
```





Out[28]:

k-value	distortion	inertias

Modeling / Clustering:

1. K-Means clustering model is used
2. Elbow method is used to find out best value of k
3. In this case best value of k is

K = 5

Clustered Australian cities