IBM Data Science Capstone Project

Picking the right location for a new Tibetan restaurant in shanghai

Kai liang

December 2020

1. Introduction

In this capstone project I will use the knowledge learned in the previous chapters to solve a practical problem and demonstrating the creation of value by applying the learned skills。This is IBM Data Science Professional certificate course on Coursera concludes with a Capstone Project. This project is about using data science toolset on a real-life problem and demonstrating the creation of value by applying the learned skills. This report presents this capstone project. The analysis was performed in Python.

2 Problem Definition

(1).problem

For this project, I chose a theoretical business problem. The question that we are trying to answer is the following.

My friend is studying in Shanghai University, he plans to start a business after graduation, therefor he decided to open a Tibetan restaurant in Shanghai,china.

Taking into account the price level at which the restaurant will operate, the intent is to find an optimal location in an area, where gastronomy is booming and which is easily accessible for tourists and for wealthier local citizens as well.

(2).business logic

we can use unsupervised machine learning to create clusters of districts that will provide us with a list of areas for consideration for the restaurant, The intent is that the restaurant to be situated close to one of the gastronomical centers and high income area

3.Data collection

To perform this analysis, we will need the following data:

1. List of the districts of shanghai

   List of districts will be obtained from wikipedia (https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai)

2. Geo-coordinates of the districts in shanghai

   Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.

3. Top venues of districts

   Top venues data will be obtained from Foursquare through an API.

4.the income data (GDP per capita in 2019) for districts of shanghai

the income data for districts of shanghai will be obtained from Shanghai Bureau of Statistics(http://tjj.sh.gov.cn)

4. Methodology

(1). Brief process

After tidying up and exploring the data, we will apply the Unsupervised machine learning technique for creating clusters of districts. We will use the silhouette score for choosing the optimal number of clusters.

(2) Data Preparation and exploration

As part of preparing the data, we start by creating a list of districts in shanghai and add the geo-coordinates of each district to a table. First I Network crawling the situation each district from wiki https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai After performing this task, we get the following table that we use in pandas dataframe format.

	Name	Chinese	Hanyu Pinyin	Division code	Area shorthand	Area (km²)	Population (2018 census)	Density (/km²)
0	Huangpu Distric	黄浦区	Huángpǔ Qū	310101	HGP	20.46	653,800	31,955
1	Xuhui District	徐汇区	Xúhuì Qū	310104	XHI	54.76	1,084,400	19,803
2	Changning District	长宁区	Chángníng Qū	310105	CNQ	38.30	694,000	18,120
3	Jing'an District	静安区	Jìng'ān Qū	310106	JAQ	36.88	1,062,800	28,818
4	Putuo District	普陀区	Pǔtuó Qū	310107	PTQ	54.83	1,281,900	23,380
5	Hongkou District	虹口区	Hóngkǒu Qū	310109	HKQ	23.48	797,000	33,944
6	Yangpu District	杨浦区	Yángpǔ Qū	310110	YPU	60.73	1,312,700	21,615
7	Pudong New Area	浦东新区	Pǔdōng Xīnqū	310115	PDX	1,210.41	5,550,200	4,585
8	Minhang District	闵行区	Mǐnháng Qū	310112	MHQ	370.75	2,543,500	6,860
9	Baoshan District	宝山区	Bǎoshān Qū	310113	BAO	270.99	2,042,300	7,536
10	Jiading District	嘉定区	Jiādìng Qū	310114	JDG	464.20	1,588,900	3,423
11	Jinshan District	金山区	Jīnshān Qū	310116	JSH	586.05	805,000	1,374
12	Songjiang District	松江区	Sōngjiāng Qū	310117	SOJ	605.64	1,762,200	2,910
13	Qingpu District	青浦区	Qīngpǔ Qū	310118	QPU	670.14	1,219,100	1,819
14	Fengxian District	奉贤区	Fèngxián Qū	310120	FXI	687.39	1,152,000	1,676
15	Chongming District	崇明区	Chóngmíng Qū	310151	CMG	1,185.49	688,100	580

I got the situation about

Name	Chinese	Hanyu Pinyin	Division code	Area shorthand	Area (km²)	Population (2018 census)	Density (/km²)

In each districts, These indicators are very useful for my analysis. Then we add in the per capita GDP data for each district in Shanghai I got table like this:

	Name	Chinese	Hanyu Pinyin	Division code	Area shorthand	Area (km²)	Population (2018 census)	Density (/km²)	Real GDP per capita(Unit: RMB 10,000)
0	Huangpu Distric	黄浦区	Huángpǔ Qū	310101	HGP	20.46	653800.0	31,955	39.61
1	Xuhui District	徐汇区	Xúhuì Qū	310104	XHI	54.76	1084400.0	19,803	25.74
2	Changning District	长宁区	Chángníng Qū	310105	CNQ	38.30	694000.0	18,120	23.78
3	Jing'an District	静安区	Jìng'ān Qū	310106	JAQ	36.88	1062800.0	28,818	21.73
4	Putuo District	普陀区	Pǔtuó Qū	310107	PTQ	54.83	1281900.0	23,380	8.71
5	Hongkou District	虹口区	Hóngkǒu Qū	310109	HKQ	23.48	797000.0	33,944	14.53
6	Yangpu District	杨浦区	Yángpǔ Qū	310110	YPU	60.73	1312700.0	21,615	15.96
7	Pudong New Area	浦东新区	Pǔdōng Xīnqū	310115	PDX	1,210.41	5550200.0	4,585	22.87
8	Minhang District	闵行区	Mǐnháng Qū	310112	MHQ	370.75	2543500.0	6,860	9.89
9	Baoshan District	宝山区	Bǎoshān Qū	310113	BAO	270.99	2042300.0	7,536	7.60
10	Jiading District	嘉定区	Jiādìng Qū	310114	JDG	464.20	1588900.0	3,423	16.34
11	Jinshan District	金山区	Jīnshān Qū	310116	JSH	586.05	805000.0	1,374	13.38
12	Songjiang District	松江区	Sōngjiāng Qū	310117	SOJ	605.64	1762200.0	2,910	8.92
13	Qingpu District	青浦区	Qīngpǔ Qū	310118	QPU	670.14	1219100.0	1,819	9.46
14	Fengxian District	奉贤区	Fèngxián Qū	310120	FXI	687.39	1152000.0	1,676	10.13
15	Chongming District	崇明区	Chóngmíng Qū	310151	CMG	1,185.49	688100.0	580	5.58

A total of 16 districts are on the table, We select districts with a population of more than 500000 and a per capita GDP of more than 60000 RMB for analysis, Select the properties we need to analyze, Because Jinshan District is a pure industrial area, we exclude it, so we get the following table:

	Name	Division code	Population (2018 census)	Real GDP per capita(Unit: RMB 10,000)
0	Huangpu Distric	310101	653800.0	39.61
1	Xuhui District	310104	1084400.0	25.74
2	Changning District	310105	694000.0	23.78
3	Jing'an District	310106	1062800.0	21.73
4	Putuo District	310107	1281900.0	8.71
5	Hongkou District	310109	797000.0	14.53
6	Yangpu District	310110	1312700.0	15.96
7	Pudong New Area	310115	5550200.0	22.87
8	Minhang District	310112	2543500.0	9.89
9	Baoshan District	310113	2042300.0	7.60
10	Jiading District	310114	1588900.0	16.34
12	Songjiang District	310117	1762200.0	8.92
13	Qingpu District	310118	1219100.0	9.46
14	Fengxian District	310120	1152000.0	10.13

But some reason I can't use the geocode python library.so I got the latitude and longitude coordinates to each district of shanghai from this webside https://jingwei.supfree.net/mengzi.asp?id=820 After performing this task, we get the following table that we use in pandas dataframe format.

	Name	Division code	Population (2018 census)	Real GDP per capita(Unit: RMB 10,000)	Latitude	Longitude
0	Huangpu Distric	310101	653800.0	39.61	31.23	121.48
1	Xuhui District	310104	1084400.0	25.74	31.18	121.43
2	Changning District	310105	694000.0	23.78	31.22	121.42
3	Jing'an District	310106	1062800.0	21.73	31.23	121.45
4	Putuo District	310107	1281900.0	8.71	31.40	121.25
5	Hongkou District	310109	797000.0	14.53	31.27	121.50
6	Yangpu District	310110	1312700.0	15.96	31.27	121.52
7	Pudong New Area	310115	5550200.0	22.87	31.22	121.53
8	Minhang District	310112	2543500.0	9.89	31.12	121.38
9	Baoshan District	310113	2042300.0	7.60	31.40	121.48
10	Jiading District	310114	1588900.0	16.34	31.38	121.27
12	Songjiang District	310117	1762200.0	8.92	31.02	121.22
13	Qingpu District	310118	1219100.0	9.46	31.15	121.12
14	Fengxian District	310120	1152000.0	10.13	30.92	121.47

In the next step of the analysis, the districts were explored in greater detail. It means venues were collected for each district via

Foursquare API. The data from Foursquare is received in json format. After arranging the data, we have up to 100 venues for each district. Venues are collected within a radius of 1000 meters from the point of district coordinates. The collected and arranged data looks like this. The following table shows some venues from the first district.

	District name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Baoshan District	Fast Food Restaurant	Office	Chinese Restaurant	Shopping Mall	Vietnamese Restaurant	Gym	Electronics Store	Furniture / Home Store	Grocery Store	Gym / Fitness Center
1	Changning District	Chinese Restaurant	Coffee Shop	Hotel	Beer Bar	Lounge	Korean Restaurant	Café	Pizza Place	Noodle House	Convenience Store
2	Fengxian District	Asian Restaurant	Coffee Shop	Steakhouse	Gym / Fitness Center	Electronics Store	Fast Food Restaurant	Furniture / Home Store	Grocery Store	Gym	Historic Site
3	Hongkou District	Multiplex	Plaza	Park	Pizza Place	Hotel	Hostel	History Museum	Historic Site	Gym / Fitness Center	Hotel Bar
4	Huangpu Distric	Chinese Restaurant	Coffee Shop	Hotel	Hotpot Restaurant	Gym	Bar	Italian Restaurant	Sandwich Place	Hostel	Movie Theater

We can check how many venues have been collected for each district. The following table gives that summary.

	District name	Division code	Population (2018 census)	Real GDP per capita(Unit: RMB 10,000)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Huangpu Distric	310101	653800.0	39.61	31.23	121.48	0	Chinese Restaurant	Coffee Shop	Hotel	Hotpot Restaurant	Gym	Bar
1	Xuhui District	310104	1084400.0	25.74	31.18	121.43	0	Hotel	Shanghai Restaurant	Fast Food Restaurant	Coffee Shop	Furniture / Home Store	Motel
2	Changning District	310105	694000.0	23.78	31.22	121.42	0	Chinese Restaurant	Coffee Shop	Hotel	Beer Bar	Lounge	Korean Restaurant
3	Jing'an District	310106	1062800.0	21.73	31.23	121.45	0	Coffee Shop	Gym / Fitness Center	Theater	Hotel Bar	Japanese Restaurant	Shanghai Restaurant
4	Putuo District	310107	1281900.0	8.71	31.40	121.25	0	Asian Restaurant	Bar	Pizza Place	Fast Food Restaurant	Japanese Restaurant	Hotel Bar
5	Hongkou District	310109	797000.0	14.53	31.27	121.50	0	Multiplex	Plaza	Park	Pizza Place	Hotel	Hostel
6	Yangpu District	310110	1312700.0	15.96	31.27	121.52	2	Chinese Restaurant	Museum	Vietnamese Restaurant	Gym / Fitness Center	Electronics Store	Fast Food Restaurant
7	Pudong New Area	310115	5550200.0	22.87	31.22	121.53	0	Coffee Shop	Korean Restaurant	Sushi Restaurant	Vietnamese Restaurant	Gym / Fitness Center	Electronics Store
8	Minhang District	310112	2543500.0	9.89	31.12	121.38	3	Bakery	Café	Vietnamese Restaurant	Historic Site	Fast Food Restaurant	Furniture / Home

(3). Analysis

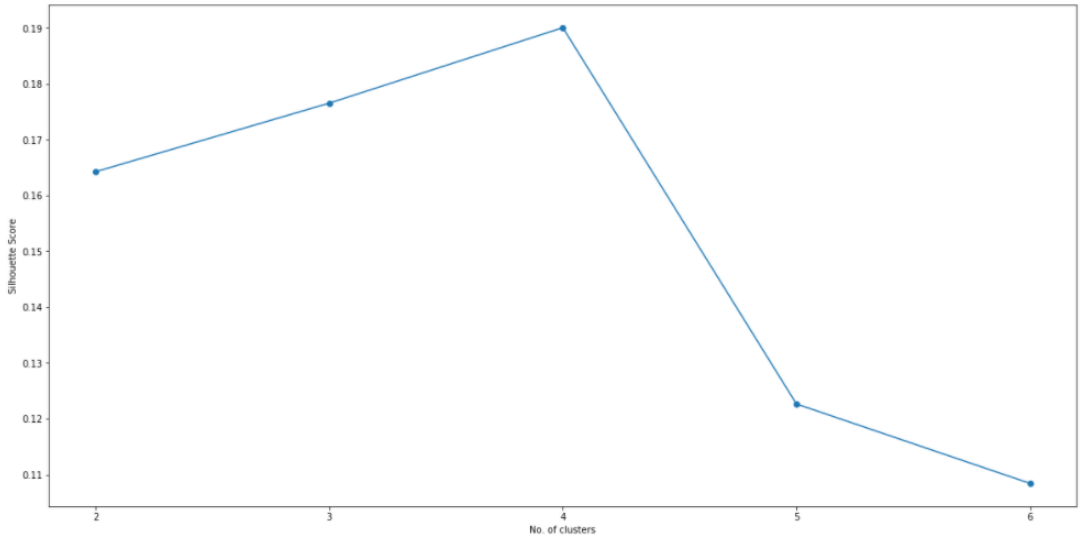
One-hot encode

For analysing the districts, we focus on venue categories. For that purpose, we use the one-hot encoding. This creates dummy variables for categories so the data set could be used for machine learning. After performing manipulations with the dataset, we get the following table, which shows the top ten most common venues for each district (first four shown in the table).

	District name	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Bar	Bistro	Boutique	Brewery	Burger Joint	...	Squash Court	Stadium	Steakhouse	Supermarket	Sushi Restaurant	Taiwanese Restaurant	T Restaur
0	Huangpu Distric	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	Huangpu Distric	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	Huangpu Distric	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	Huangpu Distric	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
4	Huangpu Distric	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
5	Huangpu Distric	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	

Clustering

Now that we have the dataset ready, we perform clustering. For this, unsupervised machine learning technique will be used based on K-means. For K-means clustering, we need to decide on the number of clusters that we want to use. To avoid the trial and error approach, the silhouette score was used. The following graph shows the silhouette scores for a range of clusters variations.



From the graph, we can read that the optimal number of clusters to use is 4 (where the score is the highest). In the next step, we run the K-means clustering algorithm with the parameter of 4 as the number of clusters. When done, we add the cluster labels to the dataset. We get the following table.

	District name	Division code	Population (2018 census)	Real GDP per capita(Unit: RMB 10,000)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th M Comn Ver
0	Huangpu Distric	310101	653800.0	39.61	31.23	121.48	0	Chinese Restaurant	Coffee Shop	Hotel	Hotpot Restaurant	Gym	Bar	Ita Restau
1	Xuhui District	310104	1084400.0	25.74	31.18	121.43	0	Hotel	Shanghai Restaurant	Fast Food Restaurant	Coffee Shop	Furniture / Home Store	Motel	Brew
2	Changning District	310105	694000.0	23.78	31.22	121.42	0	Chinese Restaurant	Coffee Shop	Hotel	Beer Bar	Lounge	Korean Restaurant	C
3	Jing'an District	310106	1062800.0	21.73	31.23	121.45	0	Coffee Shop	Gym / Fitness Center	Theater	Hotel Bar	Japanese Restaurant	Shanghai Restaurant	H
4	Putuo District	310107	1281900.0	8.71	31.40	121.25	0	Asian Restaurant	Bar	Pizza Place	Fast Food Restaurant	Japanese Restaurant	Hotel Bar	H
5	Hongkou District	310109	797000.0	14.53	31.27	121.50	0	Multiplex	Plaza	Park	Pizza Place	Hotel	Hostel	Hist Muse
6	Yangpu District	310110	1312700.0	15.96	31.27	121.52	2	Chinese Restaurant	Museum	Vietnamese Restaurant	Gym / Fitness Center	Electronics Store	Fast Food Restaurant	Furnitu Ho St
7	Pudong New Area	310115	5550200.0	22.87	31.22	121.53	0	Coffee Shop	Korean Restaurant	Sushi Restaurant	Vietnamese Restaurant	Gym / Fitness Center	Electronics Store	Fast F Restau
8	Minhang District	310112	2543500.0	9.89	31.12	121.38	3	Bakery	Café	Vietnamese Restaurant	Historic Site	Fast Food Restaurant	Furniture / Home Store	Groc St

Also, we can visualise the clusters on the map that we created earlier.

You may be surprised, but don't worry, because Shanghai is an old port city. The old city and the new city interact together. The functions of many regions are different, which affects our analysis

5. Results

Understanding the Clusters

By looking at the cluster data, we can see that cluster 2 is the one that we are the most interested in.

1. Cluster 1

The first cluster (Cluster label 0) is a Typical industrial areas where workers work , It has a large population, but its per capita income is not high, So it's not where we're going to pick

	District name	Population (2018 census)	Real GDP per capita(Unit: RMB 10,000)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
12	Songjiang District	1762200.0	8.92	31.02	121.22	1	Convenience Store	Grocery Store	Vietnamese Restaurant	Dim Sum Restaurant	Electronics Store	Fast Food Restaurant	Furniture / Home Store	Gym

2. Cluster 2

This cluster is Typical Urban Resident Area in Shanghai, It covers all aspects of residents' needs, but unlike our goal, we need a business district centered on food

	District name	Population (2018 census)	Real GDP per capita(Unit: RMB 10,000)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Cor
6	Yangpu District	1312700.0	15.96	31.27	121.52	2	Chinese Restaurant	Museum	Vietnamese Restaurant	Gym / Fitness Center	Electronics Store	Fast Food Restaurant	Furniture / Home Store	Grocery Store	
10	Jiading District	1588900.0	16.34	31.38	121.27	2	Chinese Restaurant	Bus Stop	Massage Studio	Stadium	Vietnamese Restaurant	Fast Food Restaurant	Furniture / Home Store	Grocery Store	

3. Cluster 3

The first cluster is an outer district where top gastronomy is not really represented (coffe and fast food are in the top).

	District name	Population (2018 census)	Real GDP per capita(Unit: RMB 10,000)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mo Comm Ven
8	Minhang District	2543500.0	9.89	31.12	121.38	3	Bakery	Café	Vietnamese Restaurant	Historic Site	Fast Food Restaurant	Furniture / Home Store	Grocery Store	Gym	Gyn Fitne Cent

4. Cluster 4

	District name	Population (2018 census)	per capita(Unit: RMB 10,000)	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th M Comn Ver
0	Huangpu Distric	653800.0	39.61	31.23	121.48	0	Chinese Restaurant	Coffee Shop	Hotel	Hotpot Restaurant	Gym	Bar	Italian Restaurant	Sandw Pl
1	Xuhui District	1084400.0	25.74	31.18	121.43	0	Hotel	Shanghai Restaurant	Fast Food Restaurant	Coffee Shop	Furniture / Home Store	Motel	Brewery	Squ C
2	Changning District	694000.0	23.78	31.22	121.42	0	Chinese Restaurant	Coffee Shop	Hotel	Beer Bar	Lounge	Korean Restaurant	Café	Pi Pl
3	Jing'an District	1062800.0	21.73	31.23	121.45	0	Coffee Shop	Gym / Fitness Center	Theater	Hotel Bar	Japanese Restaurant	Shanghai Restaurant	Hotel	Hot Restau
4	Putuo District	1281900.0	8.71	31.40	121.25	0	Asian Restaurant	Bar	Pizza Place	Fast Food Restaurant	Japanese Restaurant	Hotel Bar	Hotel	Ho
5	Hongkou District	797000.0	14.53	31.27	121.50	0	Multiplex	Plaza	Park	Pizza Place	Hotel	Hostel	History Museum	Hist
7	Pudong New Area	5550200.0	22.87	31.22	121.53	0	Coffee Shop	Korean Restaurant	Sushi Restaurant	Vietnamese Restaurant	Gym / Fitness Center	Electronics Store	Fast Food Restaurant	Furnitu Ho St
9	Baoshan District	2042300.0	7.60	31.40	121.48	0	Fast Food Restaurant	Office	Chinese Restaurant	Shopping Mall	Vietnamese Restaurant	Gym	Electronics Store	Furnitu Ho St
13	Qingpu District	1219100.0	9.46	31.15	121.12	0	Fast Food Restaurant	Restaurant	Hotel	Gym / Fitness Center	Dumpling Restaurant	Electronics Store	Furniture / Home Store	Groc St
14	Fengxian	1152000.0	10.13	30.92	121.47	0	Asian	Coffee	Steakhouse	Gym / Fitness	Electronics	Fast Food	Furniture / Home	Groc

Cluster 4 (Cluster label 4) is the biggest cluster, but this is where we see lots of gastronomy related venues (coffee shop, pizza place, Thai restaurant, beer bar, pub, modern European restaurant, etc..), they are the business districts,where I looking for! But how do we choose so many areas? Let's take a look at the population data and GDP per capita at the front of the table, and I'll get the answer, Huangpu District has the largest per capita income, but its population is small, But let's take a look at Pudong New Area, It has the largest population and a high per capita income. Most importantly, it is a multicultural area

7	Pudong New Area	5550200.0	22.87	31.22	121.53	0	Coffee Shop	Korean Restaurant	Sushi Restaurant	Vietnamese Restaurant	Gym / Fitness Center	Electronics Store	Fast Food Restaurant

6. Discussion and Recommendations

Based on what we learned about the clusters, we can advise the restaurant owner to consider the districts from cluster 4 as a potential location for the tibetan restaurant. These are the districts where gastronomy is well represented and also hotels are frequent. These satisfy the two original criteria that the location should be in a gastronomical centre and in a location that is easily accessible for tourists

7. Conclusion

This paper discussed the process of coming up with an answer for a hypothetical though real-life like business problem. The analysis was performed based on the toolset of data science and relied heavily on the use of Python and Python libraries such as Pandas, Scikit, Folium to name a few. Data was collected from a different type of sources and in different formats. For analysis, machine learning technique was used. The output of the analysis provided a thorough base for the recommendation for the business problem in question.

8. References

The Jupyter notebook of the analysis can be found on GitHub.

https://github.com/tibetmadman/IBM-Data-Science-Capstone-Project/blob/main/final%20data%20science%20project%20stone.ipynb