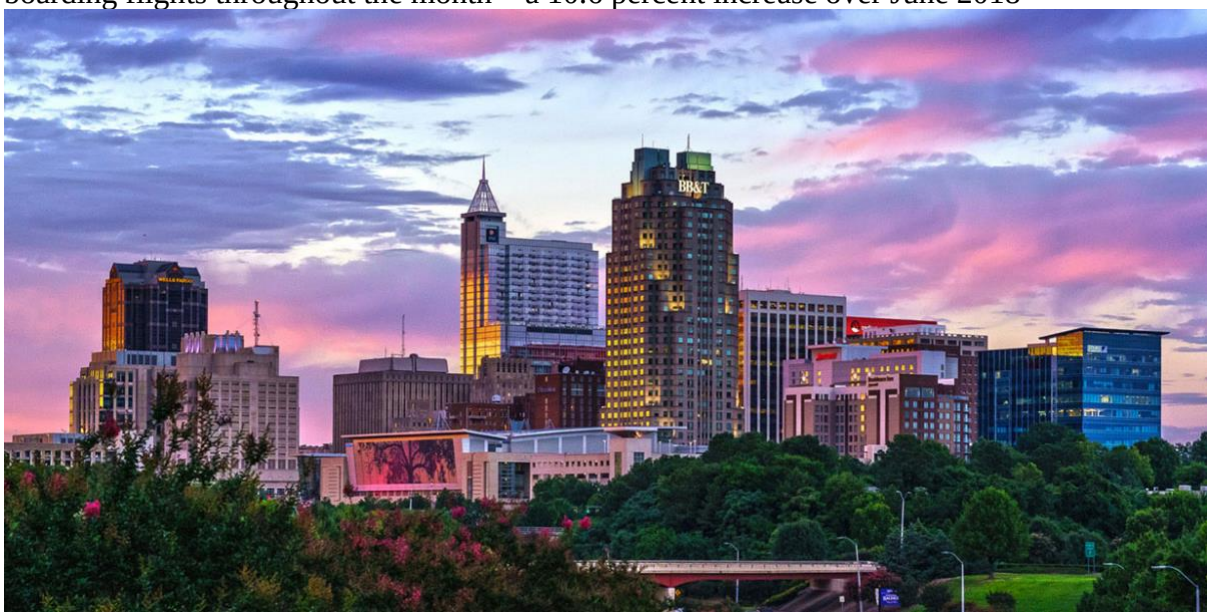
Capstone Data Science Project

By Thilak K Bhaskar
version 1.0
Sept 2019

Table of Contents

Introduction

Raleigh-Durham twin cities region also known as the RTP (Research Triangle Park) is one of the fastest growing regions of the US which is also a technology and research hub for many fortune 500 corporations including the head-quarters for RedHat, Inc. There is a huge influx and growth in population as well as a big surge in the airport traffic which indicates the amount of growth experienced by this region. The travel data for the airport traffic with Passenger volume at RDU hit an all-time high in June, with more than 679,000 travelers boarding flights throughout the month – a 10.6 percent increase over June 2018



Downtown Raleigh has emerged as a vibrant center of activity that continues to welcome historic growth and investment. As an apex of commerce and government, it is also home to a thriving creative culture with artists, musicians, innovative tech companies, award-winning chefs, and cutting-edge makers working to create the Downtown Raleigh experience.

Wake County, North Carolina's estimated population is 1,071,886 with a growth rate of 2.24% in the past year according to the most recent United States census data. Wake County, North Carolina is the 2nd largest county in North Carolina.

In this project we will analyze various data sources available to public and gather data and applied to a capstone data science project to perform data science experiments on the datasets to identify locations and provide approach and recommendations to open new restaurant business in the various suburbs of Wake County.

In this case study we have taken an example of existing locations of Mac Donald’s restaurants and applied the data to the data science tools such as Jupyter notebook using IBM Watson studio to provide the final recommendations using various data science methodologies.

Business case

As an demonstration we have taken Mac Donald's which run a big restaurant chain in the US to provide an analysis and recommendation to open a set of new restaurants chain across the wake county to serve the growing needs of the region and help them identify the best locations for launching the new restaurant outlets. This solution can be applied to any restaurant chain interested in starting new business in Wake County.

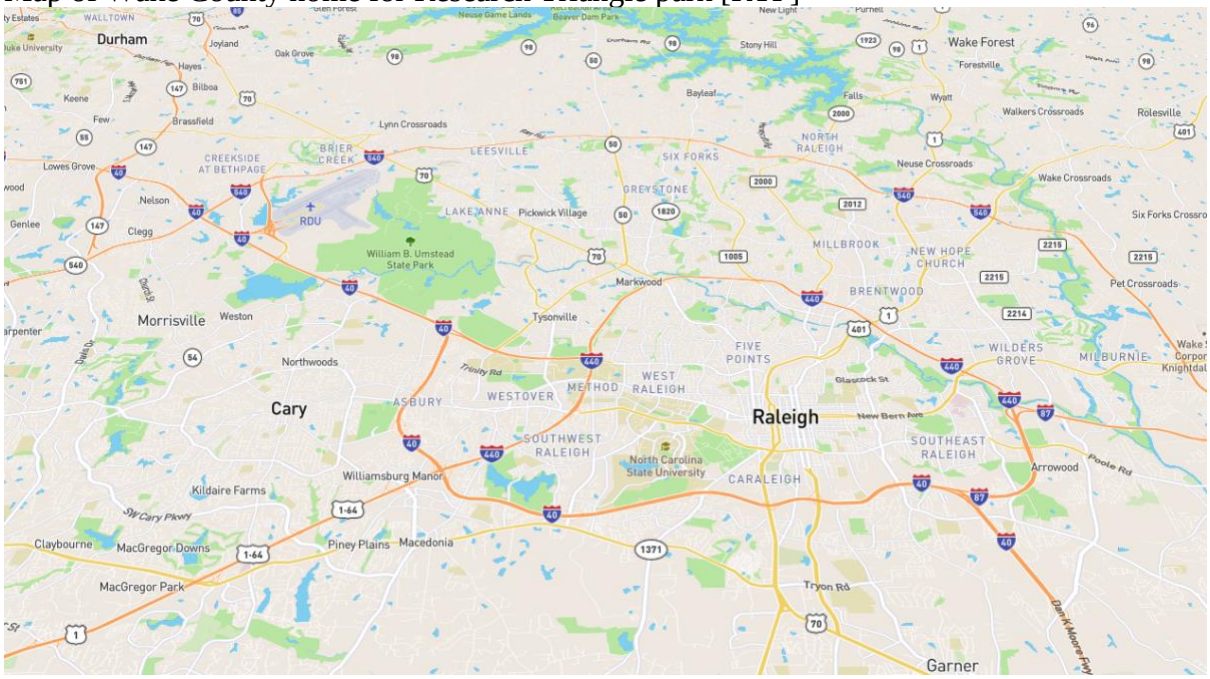
The following criteria are considered to choose the location of the new restaurant:

1. Uniqueness – This criterion checks for the availability of same chain owned restaurant in the vicinity, we were advised to identify locations which does not have the same chain restaurants within 3 miles radius.
2. Visibility and accessibility. Select a spot that can be seen by those driving or walking by. You should also look for an area that gets plenty of passersby on foot or in cars. In addition, consider if there is parking and ease of access by foot or car.
3. The demographics. Ensure the target market of your restaurant matches the demographics of the area.
4. Labor costs and minimum wage. It's important to ensure that the labor costs of an area don't cut into your profits. You will also want to have an idea of what employees might expect to make based on the location.
5. The competition of the area. Some nearby competition can help with marketing. But it's wise to have enough of a distance that you can still guarantee a solid pool of customers who won't be easily drawn to another similar place.

Data Source

Wake County – fastest growing region in North Carolina

Map of Wake County home for Research Triangle park [RTP]

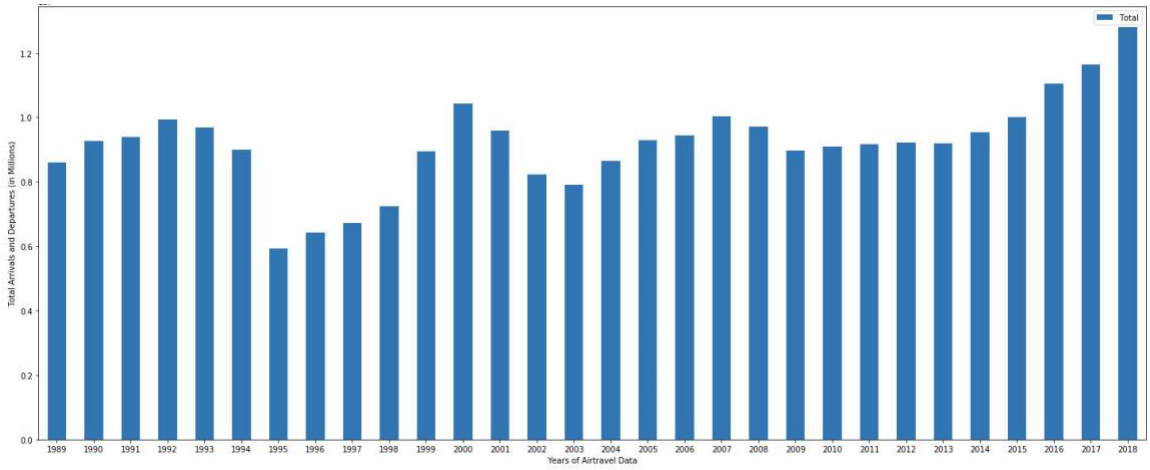


The following table gives us an Idea of growth in population in Wake County:
http://worldpopulationreview.com/us-counties/nc/wake-county-population/

Year ▾	Population	Growth	Growth Rate
2018	1,092,305	20,419	1.90%
2017	1,071,886	23,505	2.24%
2016	1,048,381	25,950	2.54%
2015	1,022,431	24,642	2.47%
2014	997,789	23,782	2.44%
2013	974,007	22,113	2.32%
2012	951,894	23,165	2.49%
2011	928,729	21,847	2.41%
2010	906,882	483,502	114.20%
1990	423,380	122,053	40.51%
1980	301,327	72,874	31.90%
1970	228,453	59,371	35.11%
1960	169,082	32,632	23.91%
1950	136,450	26,906	24.56%

RDU Data – Air traffic Data from the RDU airport portal

The following data source is also used to initially study the feasibility of opening the restaurant in the wake county based on the floating population that comes into the city by way of air traffic. Here is the sample data source captured from the RDU airport authority - *https://www.rdu.com/airport-authority/statistics/*



Restaurant Data for Wake County

In this capstone project we will address the Uniqueness requirement of the client and use the latest updated source data available in the Raleigh – Open Data portal [http://data-ral.opendata.arcgis.com/] related to existing restaurants in the wake county and provide an analysis and recommendation to identify new restaurant locations in the wake county geography using the data science approach and foursquare APIs. The following raw data was downloaded from the above link:

BUSINESS_ID	NAME	ADDRESS	CITY	STATE	POSTAL_CODE	LATITUDE	LONGITUDE
CC53271A-07F5-3B48-ADE1E62CBB5C325A	MCDONALDS #13824	101 TIMBER DR	GARNER	NC	27529.0	35.685632	-78.604503
23D6E1F8-DFC2-9B9C-55422A411A405AF7	CHAR-GRILL OF GARNER	1155 TIMBER DR E	GARNER	NC	27529.0	35.689060	-78.577039
501501	Chick-Fil-A #01488	220 Shenstone Blv	GARNER	NC	27529.0	35.692121	-78.582617
6063FF59-5056-A20B-FA9F5C4CD2B364C1	BUFFALO BROTHERS GARNER	7245 WHITE OAK RD	GARNER	NC	27529.0	35.687894	-78.580738
CC370F4E-5056-A20F-9E209C0EFB5CC41F	FAMOUS TOASTERY	52 EAGLE WING WAY	GARNER	NC	27529.0	35.697230	-78.585250
E204DECE-5056-A20B-FA9B746710617097	RED ROBIN #688	10 CABELA DR	GARNER	NC	27529.0	35.697438	-78.586124
501497	Target Food Avenue #1824	1040 E Timber DR	GARNER	NC	27529.0	NaN	NaN
A76D4311-5056-A20B-FAFAB4785D8D1BEF	BEST WESTERN PLUS EDISON INN FOODSERVICE	1595 MECHANICAL BLVD	GARNER	NC	27529.0	35.723917	-78.644718
DA48ED4A-5056-A20F-9EF54915E2942D17	Starbucks Target 1824	1040 Timber DR E	GARNER	NC	27529.0	35.687477	-78.579629
E4A60ACC-5056-A20F-9E6D9F13186DB1B5	Chhote's	1155 Timber DR E	GARNER	NC	27529.0	35.689060	-78.577039

The raw data contains 36 Cities and 3638 Restaurants.

Data Cleansing and Normalization

The raw data obtained from the multiple data sources are cleansed for the duplication and missing data and normalized to the format that can be used by the data science tools for further analysis.

	NAME	CITY	STATE	POSTAL_CODE	LATITUDE	LONGITUDE
0	MCDONALDS #13824	GARNER	NC	27529.0	35.685632	-78.604503
1	CHAR-GRILL OF GARNER	GARNER	NC	27529.0	35.689060	-78.577039
2	Chick-Fil-A #01488	GARNER	NC	27529.0	35.692121	-78.582617
3	BUFFALO BROTHERS GARNER	GARNER	NC	27529.0	35.687894	-78.580738
4	FAMOUS TOASTERY	GARNER	NC	27529.0	35.697230	-78.585250

The data frame has 36 Cities and 2871 Restaurants.

From this list of data – the filtering technique is applied using python pandas dataframes and only the existing Mac Donald's restaurant outlets are extracted

NAME	CITY	STATE	POSTAL_CODE	LATITUDE	LONGITUDE
MCDONALDS #13824	GARNER	NC	27529.0	35.685632	-78.604503
MCDONALDS #32242	HOLLY SPRINGS	NC	27540.0	35.636251	-78.831080
MCDONALD'S #32242	HOLLY SPRINGS	NC	27540.0	35.636251	-78.831080
MCDONALDS #32242	HOLLY SPRINGS	NC	27540.0	35.636251	-78.831080
MCDONALD'S #11646	KNIGHTDALE	NC	27545.0	35.798777	-78.488555
MCDONALDS AT WALMART MORRISVILLE (31892)	MORRISVILLE	NC	27560.0	35.866591	-78.846933
MCDONALDS #32336	MORRISVILLE	NC	27560.0	35.818898	-78.844822
MCDONALDS #31892 MORRISVILLE WALMART	MORRISVILLE	NC	27560.0	35.866591	-78.846933
MCDONALDS #4311	MORRISVILLE	NC	27560.0	35.818898	-78.844822
MCDONALDS #32336	MORRISVILLE	NC	27560.0	35.818898	-78.844822
MCDONALD'S CROSSROADS #13363	CARY	NC	27511.0	35.759028	-78.743340
MCDONALDS AT CROSSROADS #13363	CARY	NC	27511.0	35.759028	-78.743340
MCDONALDS #13362	CARY	NC	27511.0	35.789121	-78.829846
MCDONALD'S	CARY	NC	27511.0	35.762246	-78.782909
MCDONALDS #7501	CARY	NC	27511.0	35.762246	-78.782909
MCDONALDS #13362	CARY	NC	27511.0	35.789121	-78.829846
MCDONALD'S #27548	CARY	NC	27518.0	35.704118	-78.795636
MCDONALDS #27548	CARY	NC	27518.0	35.704118	-78.795636
MCDONALD'S #5495	CARY	NC	27519.0	35.819272	-78.901884
MCDONALDS #34433	CARY	NC	27519.0	35.819272	-78.901884

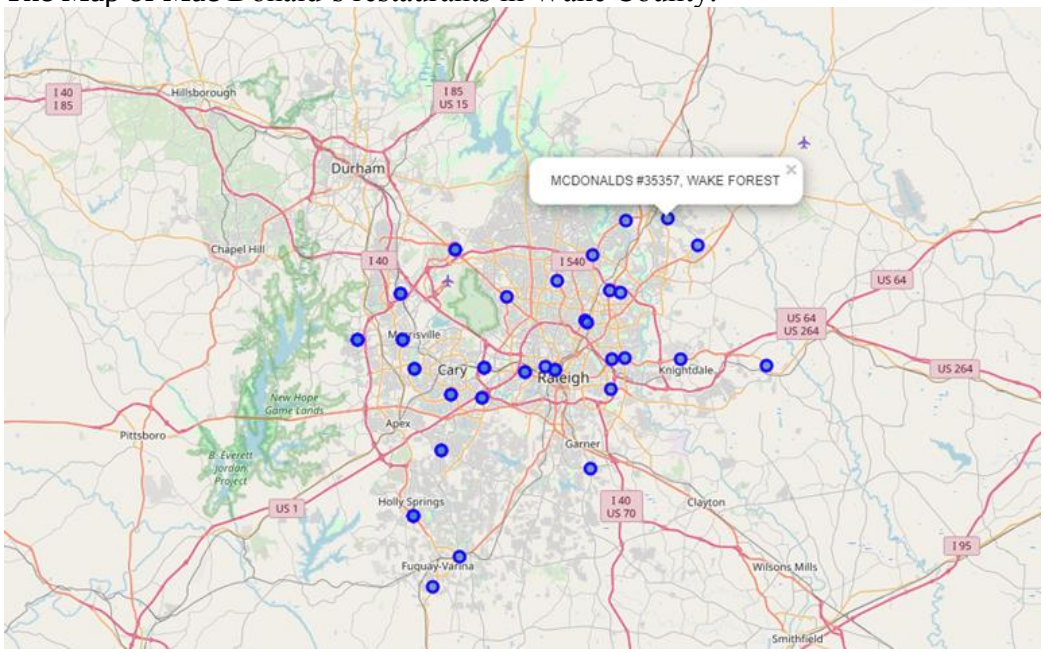
Applied Data Science Methodology

The most popular applied data science methodologies are used along with standard libraries like folium to construct the data visualization using python as a programming language
The Raleigh city air traffic data and the restaurant data which was taken from the public data site in the spreadsheet format is provided as input and the data was cleaned and normalized to be used in the python pandas data frame to be rendered in the many graphical formats and maps.

Customer segmentation is the practice of partitioning a customer base into groups of individuals that have similar characteristics. One of the algorithms that can be used for customer segmentation is K-means clustering. K-means can group data only unsupervised based on the similarity of customers to each other. Let's define this technique more formally. There are various types of clustering algorithms such as partitioning, hierarchical, or density based clustering. K-means is a type of partitioning clustering. That is, it divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm. Objects within a cluster are very similar and objects across different clusters are very different or dissimilar.

We utilized the **Foursquare API** to explore the neighborhoods and segment them. Foursquare - a location technology platform dedicated to improving how people move through the real world. Location is more than a data point. Foursquare believe that the places you go say a lot about who you are. Foursquare technology’s unparalleled sense of place and space has allowed us to help the world's leading brands and advertisers unlock valuable insights about their consumers and their businesses.

The Map of Mac Donald’s restaurants in Wake County.



Results and Data Analysis

After applying the foursquare API Calls to the existing Mac Donald's locations to find the ne arby venues within 3 kilometers radius we found that there are venues of 240 unique categori es and 3935 venues for the 10 cities

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
CARY	800	800	800	800	800	800
FUQUAY-VARINA	95	95	95	95	95	95
GARNER	73	73	73	73	73	73
HOLLY SPRINGS	213	213	213	213	213	213
KNIGHTDALE	79	79	79	79	79	79
MORRISVILLE	500	500	500	500	500	500
RALEIGH	2048	2048	2048	2048	2048	2048
Rolesville	21	21	21	21	21	21
WAKE FOREST	87	87	87	87	87	87
WENDELL	19	19	19	19	19	19

Now let's create the new dataframe and display the top 10 venues for each neighbourhood cities

	Neighborhood	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	CARY	Pizza Place	Mexican Restaurant	Supermarket	Coffee Shop	Gym / Fitness Center	Sandwich Place	Burger Joint	Fast Food Restaurant	Park	Pharmacy
1	FUQUAY VARINA	Basketball Court	Baseball Field	Bakery	Electronics Store	Soccer Field	Business Service	Miscellaneous Shop	Brewery	Boutique	City
2	FUQUAY-VARINA	Mexican Restaurant	Pizza Place	Coffee Shop	Sandwich Place	Breakfast Spot	Fast Food Restaurant	Asian Restaurant	BBQ Joint	Brewery	Bakery
3	GARNER	Fast Food Restaurant	Discount Store	Mexican Restaurant	Coffee Shop	Pharmacy	Park	Burger Joint	Japanese Restaurant	Pizza Place	Steakhouse
4	HOLLY SPRINGS	Pizza Place	Park	Sandwich Place	Fast Food Restaurant	Pharmacy	Hotel	Supermarket	Salon / Barbershop	Baseball Field	Mexican Restaurant
5	KNIGHTDALE	Pizza Place	Mobile Phone Shop	Sandwich Place	Mexican Restaurant	Fast Food Restaurant	Burger Joint	Big Box Store	Rental Car Location	Playground	Coffee Shop
6	MORRISVILLE	Hotel	Sandwich Place	Coffee Shop	American Restaurant	Indian Restaurant	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Chinese Restaurant	Convenience Store
7	RALEIGH	Fast Food Restaurant	Sandwich Place	Mexican Restaurant	Pizza Place	American Restaurant	Coffee Shop	Spa	Hotel	Grocery Store	Burger Joint
8	Rolesville	Fast Food Restaurant	Pizza Place	Sandwich Place	Hot Dog Joint	Gas Station	Pharmacy	Coffee Shop	Park	Discount Store	Seafood Restaurant
9	WAKE FOREST	American Restaurant	Fast Food Restaurant	Pizza Place	Pharmacy	Grocery Store	Italian Restaurant	Gas Station	Auto Garage	Japanese Restaurant	Supermarket

Conclusion:

From the results we gathered from the various data analysis techniques and methods we can come to the following conclusions by looking at the below tables:
The Top 5 common venue categories for the 10 cities are as follows:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	CARY	Pizza Place	Mexican Restaurant	Supermarket	Coffee Shop	Gym / Fitness Center
1	FUQUAY VARINA	Basketball Court	Baseball Field	Bakery	Electronics Store	Soccer Field
2	FUQUAY-VARINA	Mexican Restaurant	Pizza Place	Coffee Shop	Sandwich Place	Breakfast Spot
3	GARNER	Fast Food Restaurant	Discount Store	Mexican Restaurant	Coffee Shop	Pharmacy
4	HOLLY SPRINGS	Pizza Place	Park	Sandwich Place	Fast Food Restaurant	Pharmacy
5	KNIGHTDALE	Pizza Place	Mobile Phone Shop	Sandwich Place	Mexican Restaurant	Fast Food Restaurant
6	MORRISVILLE	Hotel	Sandwich Place	Coffee Shop	American Restaurant	Indian Restaurant
7	RALEIGH	Fast Food Restaurant	Sandwich Place	Mexican Restaurant	Pizza Place	American Restaurant
8	Rolesville	Fast Food Restaurant	Pizza Place	Sandwich Place	Hot Dog Joint	Gas Station
9	WAKE FOREST	American Restaurant	Fast Food Restaurant	Pizza Place	Pharmacy	Grocery Store

We can see that there are no other concentration of fast food restaurants within 3 kilometers for Morrisville, Fuquay-varina and Cary.
So we can recommend that they do further exploratory data analysis in these region to full fill other requirements for opening new restaurant outlets.

References

Four Square API - https://foursquare.com/developers/apps
IBM Hybrid cloud platform – Watson studio
Wake County Population data - http://worldpopulationreview.com/us-counties/nc/wake-county-population/
RDU Air Traffic Statistics - https://www.rdu.com/airport-authority/statistics/
Raleigh – Open Data portal - http://data-ral.opendata.arcgis.com/