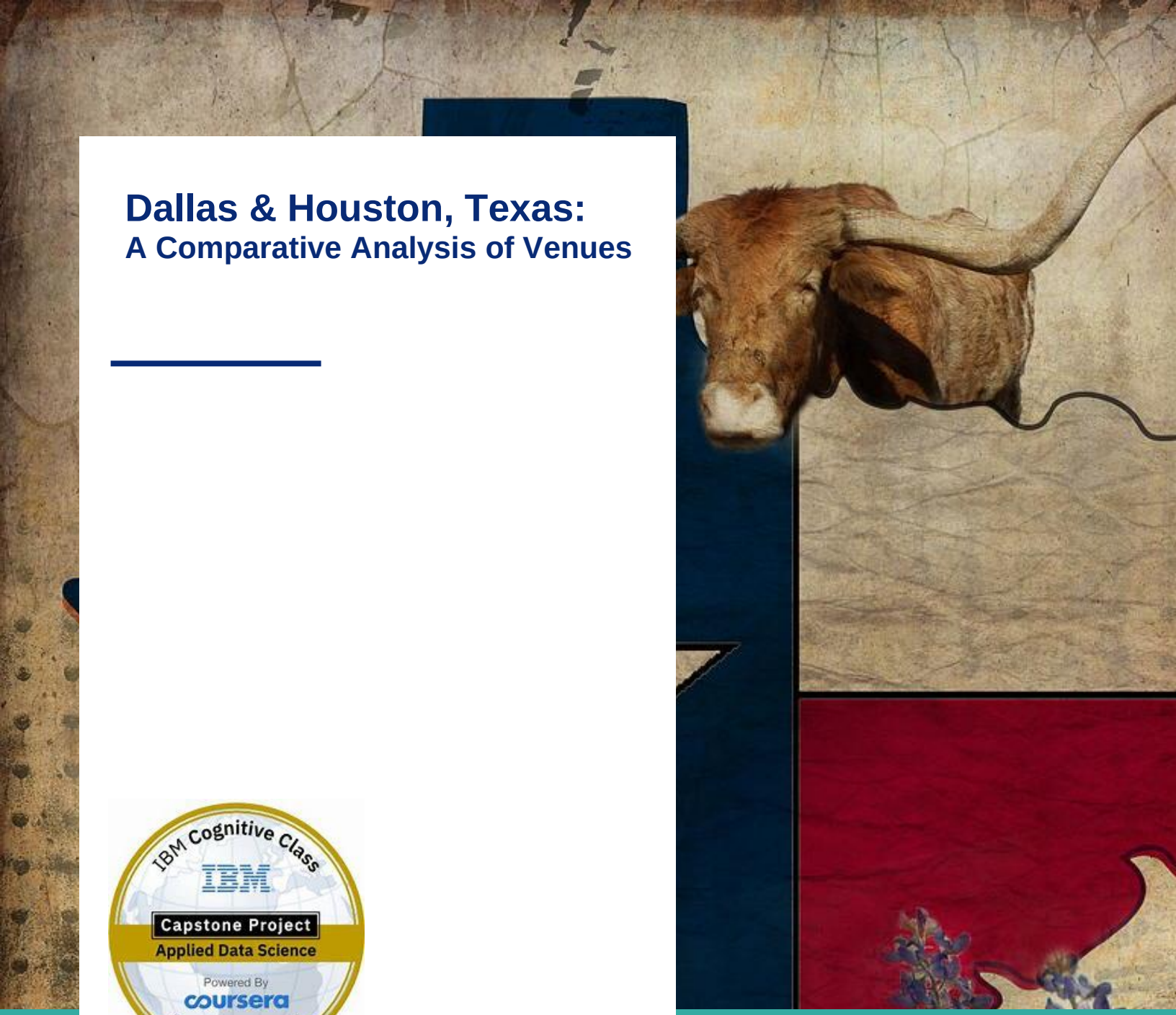

Dallas & Houston, Texas:
A Comparative Analysis of Venues

JULY 19

IBM Data Science Capstone Project
Authored by: Paul Bristow

Introduction

PROBLEM DESCRIPTION

Having lived in Texas for approximately 21 years, I wanted to better understand the differences between two of the largest and most populous cities in Texas. This report will compare and contrast the distribution of venues throughout the City of Dallas (Population = 1.34M, Area = 385.8 sq Miles) and the City of Houston, TX (Population = 2.31M, Area = 637 sq Miles). Using various data gathering, analytic and visualization tools learned throughout the Data Science Certification Program from IBM through Coursera, this report will use quantitative analysis to provide a better understanding on the distribution of venues throughout the active zip codes associated with these two cities.

The intent of this document is not to dissuade or present a bias view, but to provide a sound analytical approach to examining and documenting the distribution of venues throughout Dallas and Houston.

As a result of this report, readers considering relocating or visiting Dallas or Houston should be enabled to make better decisions based on their preferences associated to the venues and distribution of those throughout these two cities.

DATA

For this analysis, the data used will be retrieved from multiple sources, including:

- Gas Lamp Media Blog – file with all the US zip codes and their associated latitude, longitude, city, state, and county. (First 5 rows of represented in Figure 1)

	zip_code	latitude	longitude	city	state	county
0	501	40.922326	-72.637078	Holtsville	NY	Suffolk
1	544	40.922326	-72.637078	Holtsville	NY	Suffolk
2	601	18.165273	-66.722583	Adjuntas	PR	Adjuntas
3	602	18.393103	-67.180953	Aguada	PR	Aguada
4	603	18.455913	-67.145780	Aguadilla	PR	Aguadilla

Figure 1 - http://docs.gaslamp.media/wp-content/uploads/201

- World Population Review Site – file containing 2021 active Texas zip codes with their associated population

	Zip Code	City	County	Population
0	77449	Katy	Harris County	128294.0
1	77494	Katy	Fort Bend County	118291.0
2	79936	El Paso	El Paso County	111620.0
3	75034	Frisco	Collin County	108525.0
4	77084	Houston	Harris County	107673.0

Figure 2 - https://worldpopulationreview.com/zips/texas

- Data will be retrieved through Foursquare API searching for “venues” throughout the active Dallas and Houston ZIP Code regions.

Data from the first two sources will be merged to extract only active Zip codes located in each of the cities. The data from this merged file will be used with the Foursquare

API to provide segmentation, clustering and visualization of similar regions throughout Houston and Dallas.

Using an unsupervised machine learning technique with the merged data and the venues gathered from the Foursquare API, K-Means clustering will provide segmentation and clustering data. Through georeferencing and Python Folium, as well as, the graphical Python plotting library, MatPlotLib, maps and charts will be used to visualize regions and the distribution of venues throughout Dallas and Houston.