2020

Keeneland Venue Analysis and Investment Recommendations

IBM DATA SCEINCE CAPSTONE PROJECT

ANDREW BAKERT

Executive Summary

Keeneland Race Course is a central component of the economy and the culture of Lexington, KY, and the Bluegrass region. The 2019 fall meet alone attracted 262,630 attendees (Keeneland, 2019). In November 2020 the Breeders' Cup will return to Keeneland. This two-day event is one of the largest horse racing events in the county and in the world. The last time Keeneland hosted the event in 2015 it attracted 95,102 attendees (Breeders' Cup, 2015).

The large impact of Keeneland on the local economy owing to these high attendance numbers presents an opportunity for new investment in venues catering to attendees. While a study at the University of Kentucky (Bollinger, 2015) estimated the impact by sector, the specific impact by category of venue and which areas of the city are most impacted have yet to be determined. This study aims to bridge these gaps using the Foursquare Places API (Foursquare, 2020).

In order to quantify the number of visitors by category and location, an evaluation method was devised based on which venues were visited after Keeneland with six layers of analysis conducted. This yielded the number of visits by venue which was used to determine the most popular categories. Points were then created based on locations visited and the proximity of visits was determined. The top categories not within 1 km of these points were determined and these categories were used to find which categories not visited by Keeneland goers are nearby to the points.

Four metrics were then used to evaluate the relative value of each location-category pair, and inherently inferior pairs were excluded. Resulting from this analysis are seven location-category pairs shown in Table 1. Each pair has strengths and weaknesses and the final decision of which site to develop will depend on the relative importance of attracting Keeneland goers versus the general population.

Table 1: Location-category pairs determined to be most ideal

Category	**Minimum Distance (m)**	**Venues in 1 km radius**	**Visits in 1 km Radius**	**Visits in Category**	**Address**
Ice Cream Shop	2,000	-	8,600	19,000	230 Plaza Dr
Ice Cream Shop	2,000	-	8,600	19,000	278 Southland Dr
Liquor Store	2,000	-	97,000	12,000	3801 Harrodsburg Rd
Liquor Store	2,000	-	97,000	12,000	3765 Palomar Centre Dr
Steakhouse	1,123	1	6,800	46,000	455 Southland Dr
Steakhouse	839	1	24,000	46,000	3191 Beaumont Centre Cir
Steakhouse	2,000	-	320	46,000	101 Virginia Avenue

Table of Contents

List of Figures/Equations/Tables

Section 1. Introduction

Many people associate horse racing and breeding with the state of Kentucky, and Central Kentucky in particular. Lexington, Kentucky's official slogan is "Horse Capital of the World," and it's impossible to travel backroads in the Bluegrass without driving by miles and miles of pristine horse farms. Although Lexington's economy has expanded beyond horse racing alone, owing to a large university and decades of relatively rapid development, the city is still highly dependent on the industry and has established an urban/rural boundary to prevent over-development of farmland surrounding the city.

The center of horse racing in the city is the racetrack Keeneland. Hosting 18 stakes races in its fall 2019 meet with six grade I races and purses topping out at $1 million (Staff, 2019), the track attracts spectators both from the US and globally. The 2019 fall meet had attendance of 262,630 people (Keeneland, 2019). Considering the size of the city is around 330,000 people (US Census Bureau, 2019), the track drives a large part of the local economy during this period. Another meet is scheduled every year in April, with the 2019 spring meet having an attendance of 242,547 people (Keeneland, 2019). The 2020 Breeders' Cup, one of the premier racing events in the US and in the world, is also on the horizon in November 2020. When Keeneland last hosted this two-day event in 2015 it attracted record attendance with 95,102 attendees (Breeders' Cup, 2015).

Because Keeneland is very important to the city and clearly attracts a great number of visitors, the indirect benefits of racing in the area could influence decision making and new development. As such, the central problem proposed is how to quantify the indirect impact of Keeneland on different neighborhoods and venues. Also, how can this information be leveraged by someone looking to start a business in the area that caters to racetrack attendees? Developing a business themed around horse racing could be a promising venture if placed in the correct location and in the right category. Ideally this business would be relatively easy to start without requiring a great deal of capital investment.

Section 2. Data

The following data sources were used in determining the ideal location and category placement:

- The Foursquare API (Foursquare, 2020) was used to determine the next venues that users visit after checking in to a prior venue. This process was needed to evaluate the impact of Keeneland's visitors on the local economy. The API was also used to determine what venues are nearby to each chosen location and category. This was necessary to determine two metrics for determination of ideal locations and venue categories: number of venues of that category within a 1 km radius and minimum distance to a venue of that category.

- A shapefile for the state of Kentucky (US Census Bureau, 2019) provided by the census bureau was used to determine the location of each studied point by tract.
- A shapefile for zoning area for the city of Lexington (LFUCG, 2020) was obtained to filter possible locations to include only those located in business zones
- The Google Maps Geocoding API (Google, 2020) was used to determine the addresses of each chosen location based on the location's coordinates.

Section 3. Literature Review

To analyze the economic impact of Keeneland and racing on the local economy a study was conducted at the University of Kentucky. This study (Bollinger, 2015) measured the economic impact in two areas: racing and sales. In racing, the author relied on survey and sales data to determine that the top categories of spending annually were in retail (over $12 million), lodging (over $15 million), food and beverages (over $19 million), and gasoline ($9 million). The total annual economic impact from racing was estimated at $123 directly and $200 million total. This study is the first to estimate Keeneland's economic impact.

Although this study measured the economic impact of racing in the area generally, it did not analyze which areas of the city were most impacted and how more specific categories were impacted. This study aims to further evaluate these areas and to recommend specific locations and categories of venues for investment. Based on the prior study, food and beverage may be the best categories for investment, and locations in the city with close proximity to Keeneland may be ideal for development.

Section 4. Methodology

4.1. Determination of next venues

4.1.1. Venues

To approach the problem of developing a business themed around horse racing first requires an understanding of which areas are most visited by attendees and which categories of venue are most frequented. The Foursquare API (Foursquare, 2020) includes an option to search for "Next Venues". This search requires a venue id and returns the top five, or fewer if not available, venues that users who checked into the searched venue have checked into next. Using Keeneland as level 0, the next venues search was used to determine where visitors will check in after the prior venue.

4.1.2. Weights

In addition to evaluating the next venues at which visitors checked in, this analysis created a weighing scheme to determine how many visitors checked in at each venue. As described in the Foursquare documentation, the next venues are listed in descending order based on number of check-ins after the prior venue. In order to determine the next weights, a linear relationship between weights was assumed. As such, the weighing function in Equation 1 was used to determine each next venue's weight, where "n" is the number of venues and "i" is "n" minus the index of the venue.

Equation 1: Next venue weight

$$weight_i = i \times \frac{1}{(n+1) \times (n//2) + (n \,\%\, 2) \times \frac{n+1}{2}}$$

After determining each venue's next weight, each of the weights was multiplied by the prior weight to obtain the new weight to be used in the next level. For example, the weight of Keeneland at level 0 was 1, the weight of its first next venue in level 1 was $1 \times \frac{5}{15} = 0.33$, and the weight of the first next venue for this venue in level 2 was $0.33 \times \frac{5}{15} = 0.11$. This process was performed for six levels of analysis.

4.1.3. Tract Number

While computing the weights for each level, the process relies on census tract data to determine the census tract for each venue. A shapefile provided by the US Census Bureau (US Census Bureau, 2019) was used to match each venue with its corresponding tract. This shapefile was first converted to a geoJSON file using the Geopandas library. The geoJSON file was used along with the latitude and longitude of each tract to determine the tract to which each venue belongs. This determination was based on the geometric definition of a polygon and the linear relationship between each set of points in the coordinates given by the geoJSON. As the focus of the study is solely the city of Lexington, the tracts were filtered to include only Fayette County, which has a county code of "067". Additionally, the rural tracts that Keeneland visitors impacted were not included in those studied.

4.2. Creation of points of interest to study

4.2.1. Grouping and clustering venues data

After determining the proportion of the original visitors that each venue has at each level, those levels were first concatenated and then grouped by venue to determine the total impact. The total number of visitors was determined by multiplying by the number of visitors in the fall 2019

meet, 262,630 (Keeneland, 2019). After a data frame was created with the total number of visitors per venue, the venues were clustered together using a K Means algorithm. The algorithm was fitted with number of clusters varied from 1 to 14. Shown in Figure 1 is an elbow chart of the resulting sum of squared distances from each choice of number of clusters. From this chart, the number of clusters was chosen as 4.

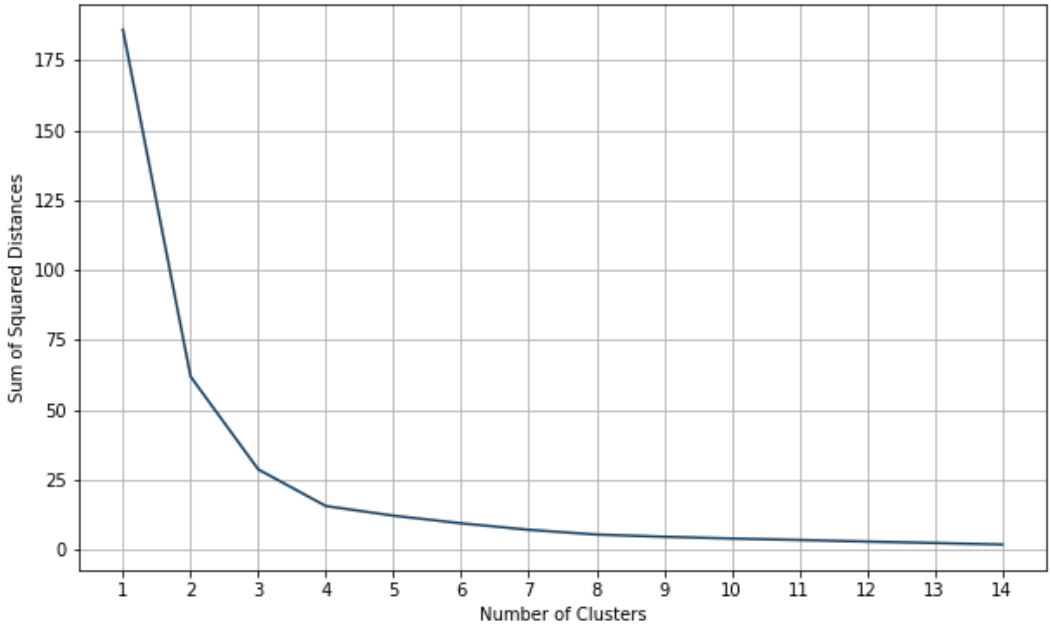

Figure 1: Elbow chart for K Means analysis of Lexington venues affected by Keeneland

Shown in Figure 2 is a heatmap of all venues with visitors from Keeneland with the chosen clusters and the 1 km radius area surrounding the clusters overlaid. The heatmap measures density of venues, so it does not include total number of visits. A map was created to illustrate the number of visitors by tract. However, this was not included in the final notebook or report as the visits were very concentrated in a few areas. The most visited urban area, census tract 21067004208, is represented by the cluster to the furthest west. This tract had roughly 120,000 visits based on the analysis conducted. The next most visited urban tract only had 5,000 visits. Proximity to Keeneland, labelled in green on the middle west side of the map, was likely to account for the high number of visits.

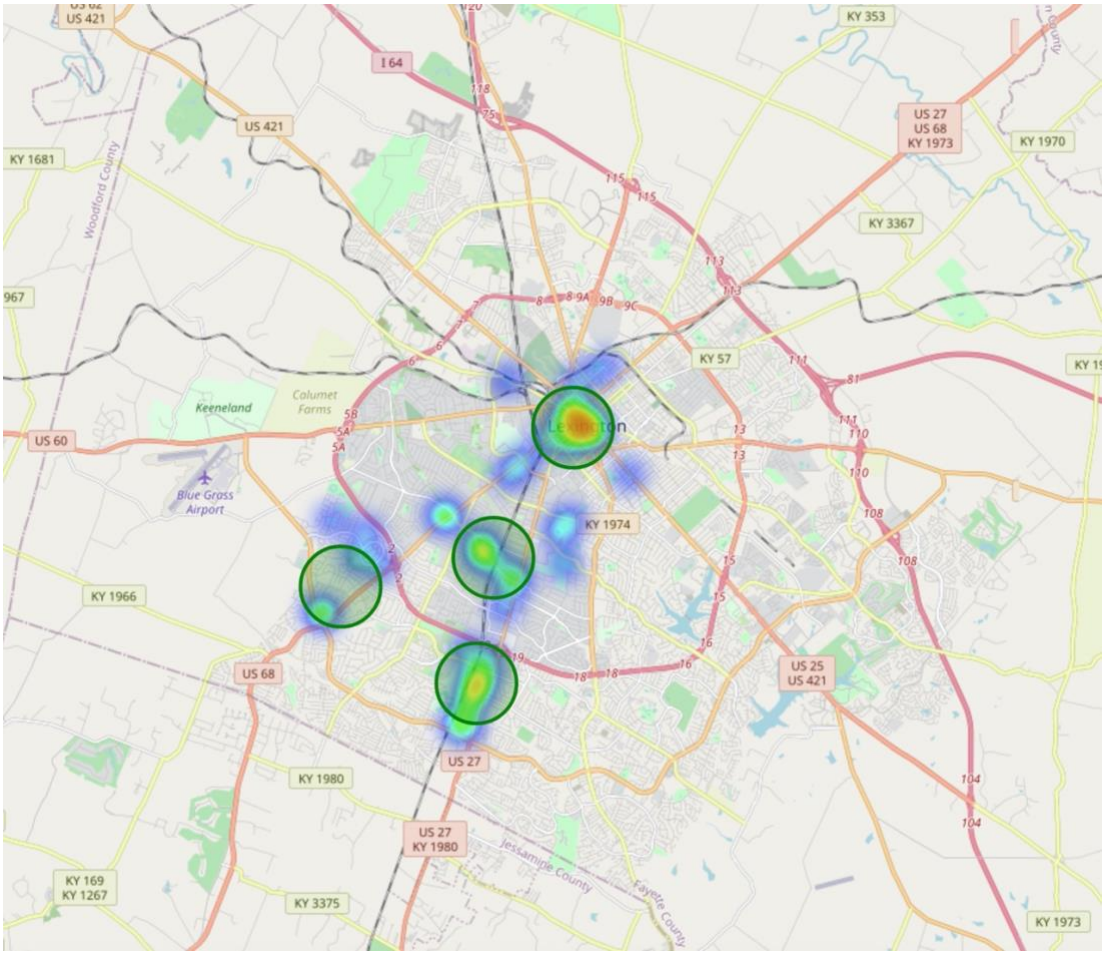

Figure 2: Venue location heatmap with k means cluster overlay

4.2.2. Studied points generation

A grid of points was then constructed based on the cluster centers. In order to create this grid, the UTM package was first used to convert latitude and longitude coordinates to x and y. An 11 x 11 grid was constructed using the cluster centers as the centers of the grid and with five layers of points chosen both vertically and horizontally separated from the center and 300 m apart. The resulting grid thus spanned 3 km vertically and horizontally.

Using the zoning shapefile for Lexington (LFUCG, 2020), which was converted to a geoJSON file, the points were filtered according to their zones. The schedule of zones for the city (LFUCG Planning Office, 2020) includes several zones for businesses. All zones starting with “B”, including “B-1”, “B-2”, “B-2A”, “B-2B”, and “B-3” were included, with the exception of “B-4”, which only applies to wholesale and represents businesses that would be too large for consideration in this study. The zones for each point were determined using the same geometric definition of points inside of a polygon as used for determining tracts. The points were then filtered to exclude all points not in an included zone.

Shown in Figure 3 is the clustered map of the locations of studied points that are within the business zones. Most of the points are in the downtown area of Lexington as well as along Nicholasville Road, marked as “US 27”, and Harrodsburg Road, marked as “US 68”.

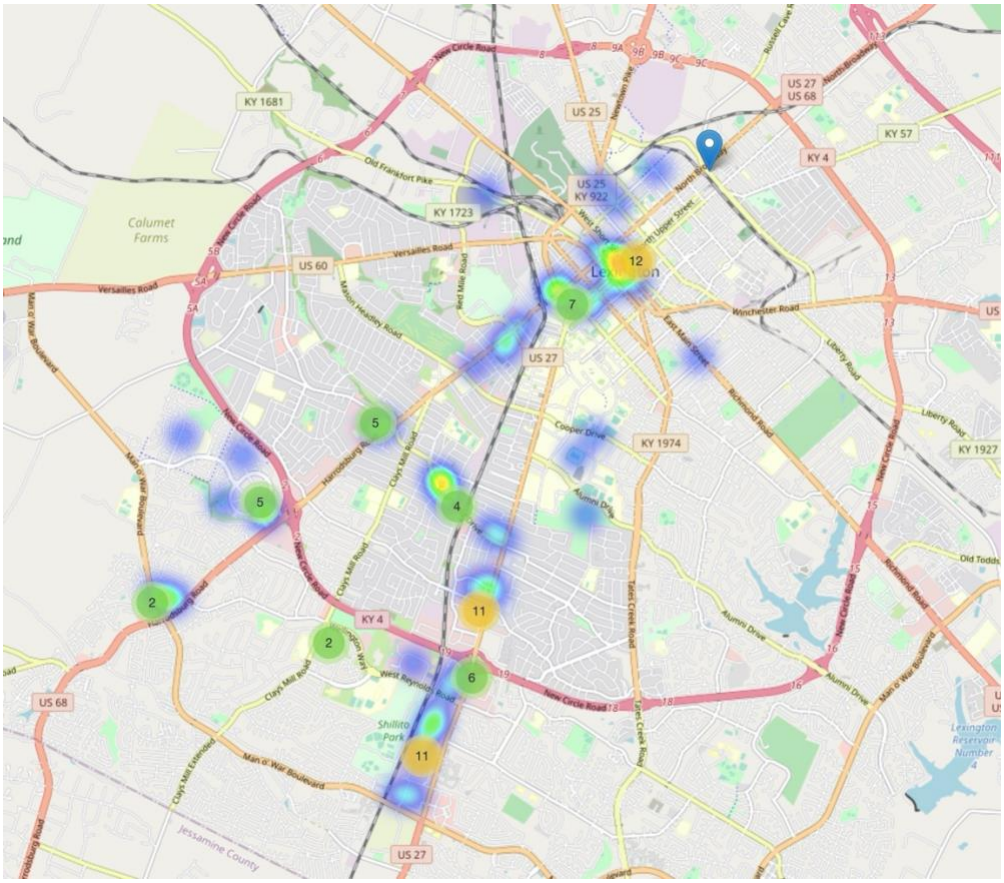

Figure 3: Clusters of chosen points for inclusion in the study

4.3. Category choice by point

4.3.1. Determining categories at each point

Next the distance to each venue visited by Keeneland goers was calculated for each point. This was determined using the data frame containing all venues with Keeneland visitors and a distance calculation function, again using the UTM package to transform latitude and longitude. After all points were matched to their distance to all venues, a radius of 1 km was chosen to filter venues and obtain only nearby venues for each point. The points and their nearby venues were then grouped by category to obtain the categories nearby and the total number of visitors per category.

4.3.2. Top categories determination and point category selection

To determine which of these groups are most visited, the venue data frame was then grouped by category and the number of visitors for each category was summed. Sorting in descending order by visits yielded the ranking of categories by visitors. The top categories were chosen based on total visitors. Supermarkets and grocery stores were excluded from the choice of top categories because they were deemed too large of an investment to be in the scope of this analysis.

The top 15 categories by total visits are shown in Figure 4. The 5 categories chosen are shown in red. Steakhouses are the most visited category with roughly 46,000 visits, with ice cream shops attracting less than half that number of visits at roughly 19,000. The number of visits per category also decreased substantially after supermarkets, with supermarkets attracting 9,000 visits and German restaurants attracting 4,000 visits.

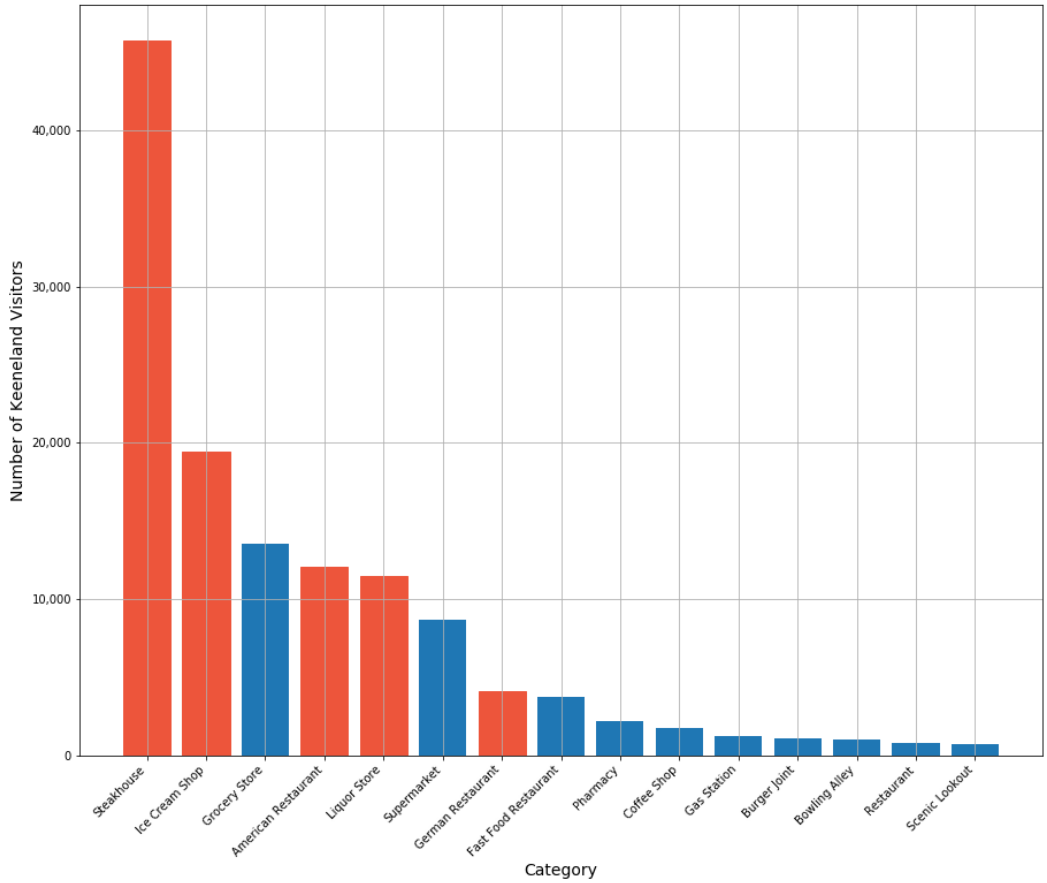

Figure 4: Top 15 categories of venue based on Keeneland visits

Next, the top categories were matched to the categories within a 1 km radius of each point to obtain a list of nearby most-visited categories. For each point the top categories that were not visited within a 1 km radius were determined by matching the existing top categories nearby with the list of all 5 top categories. The top categories not located within a 1 km radius of each point

were then matched up with category ID using a Foursquare API call. This API call yielded the category names and ID numbers for all categories delineated within the Foursquare database.

4.4. Nearby venue and location metric derivation

4.4.1. Nearby venue determination

The nearby venues to each point by category were then determined, again using the Foursquare API. The category IDs for each category and the latitude and longitude of each point were used within the API, along with a limit of 20 venues and a radius of 1 km, to determine the nearby venues. The name of each venue and its distance from the point were extracted from the results. For some areas, there were no venues for a chosen category with a 1 km radius. For each of these locations, the name was assigned a null value, while the distance was set to 2 km to indicate a greater distance from the point than within the scope of the search.

4.4.2. Venue metric calculation and justification

Next the points and their nearby venues by category were grouped by coordinates of each point and the categories. Using this grouping the number of venues in each category within a 1 km radius and the minimum distance to each category were determined for each point. As the categories with no nearby venues were assigned null values, the number of venues in each category within 1 km was 0. Although the radius was set at 1 km, some of the venues returned were slightly more than 1 km away from the pairs considered.

In addition to the number of nearby venues in each category and the minimum distance to a venue in each category, the number of visits within 1 km of the point and the number of visits per category were included. The data frame including all of the chosen points and existing categories with high visits was grouped by location and category, and visitors were summed for the location from every category. This calculated the number of visits within 1 km of the point. The number of visits per category was taken from the grouped category data frame used to determine the top categories.

Two metrics, distance to nearest venue of each top category and number of venues in that category within 1 km, characterize the merit of each category and location pair based on all existing venues. Two additional metrics, number of visits within 1 km of the point and number of visits in the studied category, characterize the merit of each pair based on their popularity among Keeneland visitors.

4.5. Location-category pair distribution

4.5.1. Distribution by metric

In order to get a better understanding of the distribution of the four chosen metrics across all studied points, histogram subplots were generated for each metric. Each histogram was grouped using 30 bins, except for in category visits, which was grouped using 20 bins.

As shown in Figure 5, the minimum distance is skewed left with many of the points with nearby venues having a relatively short distance to those venues. However, there are a large number of venues with a 2 km minimum distance, which, as previously described, indicates that there were no nearby venues in that category within 1 km. The histogram for number of venues nearby indicates that again the distribution is skewed left, with most location-category pairs having under 4 venues nearby. There is a sizable number of pairs with 0 nearby venues.

Number of visits in a 1 km radius is heavily skewed left, with the vast majority of points having relatively very few visits. However, there is one area with vastly more visits, at around 97,000. This indicates that the number of visitors from Keeneland is highly concentrated in certain areas but relatively low at most locations. The number of visits in each category is again slightly left skewed, but relatively uniform. There is a sizeable gap in visits, as can be seen in Figure 4, between the top category, steakhouses, and other categories.

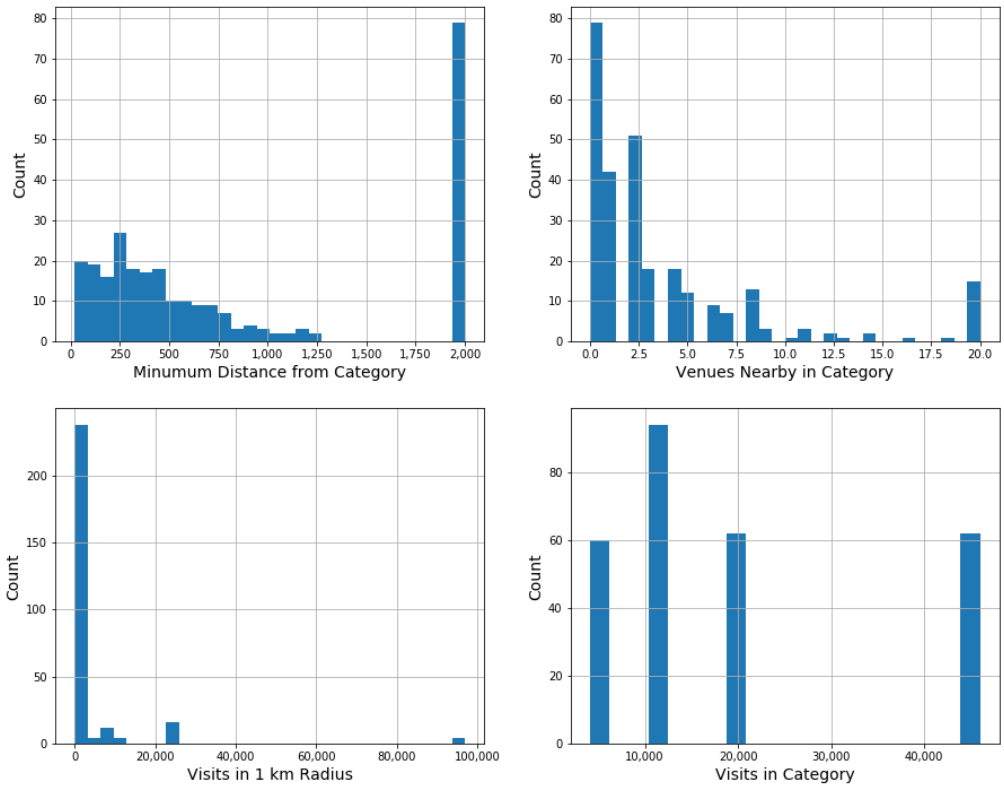

Figure 5: Distribution of evaluated metrics for all considered points

4.5.2. Distribution by category and metric

A series of subplots was then used to further examine the distributions by category, shown in Figure 6. These distributions are shown as boxplots for minimum distance from category, venues nearby, and visits in a 1 km radius. The outliers are excluded for this visualization. Because there are only 5 categories chosen, the visits by category is discrete and is shown again as a bar plot. American restaurants have the narrowest distribution for minimum distance with the lowest values for quartiles and median. This category also has the highest values for quartiles and median and the widest range for nearby venues. This indicates that American restaurants are widespread and numerous.

Ice cream shops have the greatest range for minimum distance and the highest median and third quartile for number of venues nearby. These venues also appear to be nearby to a large number of pairs, while their distance from the pairs varies widely. Steakhouses appear to be the least common in nearby venues, with by far the lowest third quartile. While the medians and ranges differ slightly, the number of visits within a 1 km radius seem to be fairly similarly distributed across all five categories. No single category is superior across all four metrics based on distribution.

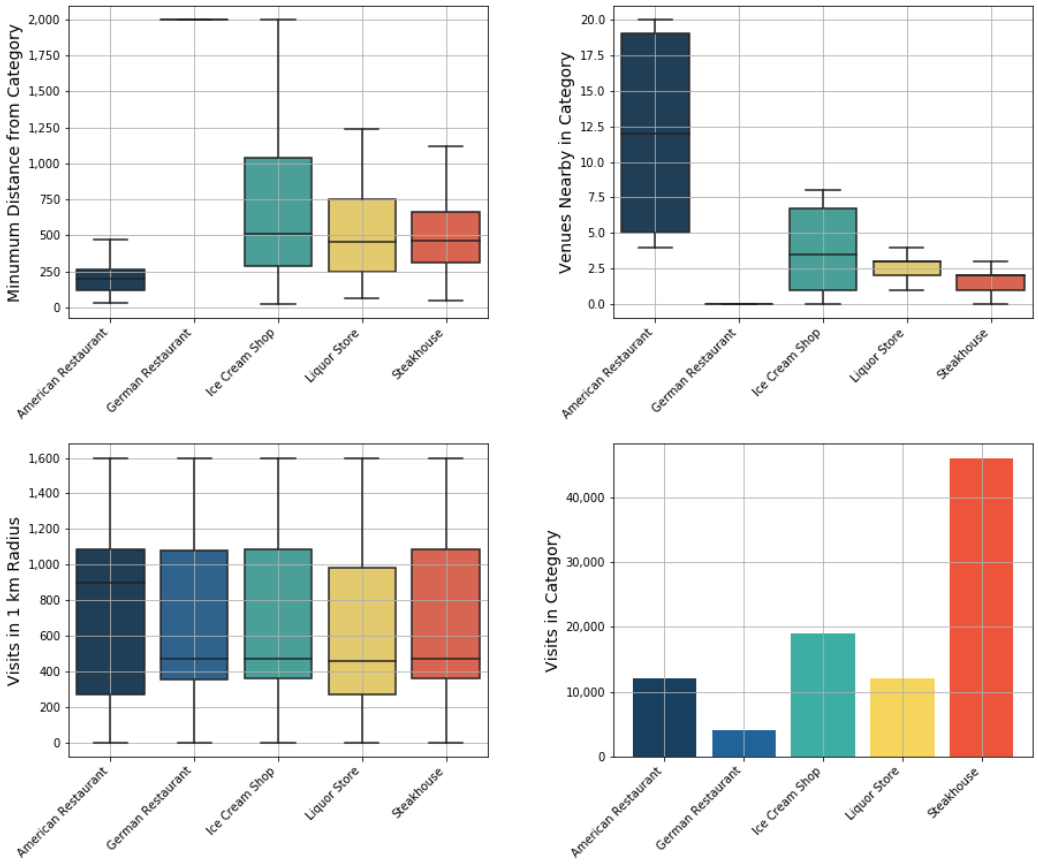

Figure 6: Distribution of metrics across all five top categories

4.6. Final location-category pairs choice

After the distribution of location-category pairs was examined both by metric and by metric and category, the ideal location-category pairs were then chosen on the basis of these four criteria. In order to determine which pairs are ideal, a function was used to filter out pairs that were inherently inferior based on the four metrics. For a pair to be inherently inferior at least one other pair must be superior by at least one metric. All other metrics need to be either equal or inferior than those of the other pair. This determination of inferiority or superiority significantly reduced the number of pairs considered and yielded the final recommended categories and locations.

The addresses of these pairs were then determined to provide more specific and useful information for potential clients. The Google Maps Geocoding API (Google, 2020) was used to match the latitude and longitude coordinates with specific addresses in Lexington. These addresses were then added to the data frame including chosen location-category pairs.

Section 5. Results

5.1. Metrics and categories for chosen location-category pairs

Shown in Table 2 are the results from the analysis and the final recommended location-category pairs. The metrics for each pair are also included to enable clients to weigh the relative importance of each metric. As seen in Table 2 the pairs chosen span three categories: ice cream shop, liquor store, and steakhouse. All but the steakhouse location on Southland Drive and the steakhouse location on Beaumont Centre Circle have no venues in the same category located within 1 km. All pairs had a high number of visits by category and all but the steakhouse on Virginia Avenue had more than 5,000 visits within a 1 km radius. Both ice cream shop locations are located in close proximity to each other, as are the liquor stores locations to each other.

Table 2: Final chosen location-category pairs

Category	**Minimum Distance (m)**	**Venues in 1 km radius**	**Visits in 1 km Radius**	**Visits in Category**	**Address**
Ice Cream Shop	2,000	-	8,600	19,000	230 Plaza Dr
Ice Cream Shop	2,000	-	8,600	19,000	278 Southland Dr
Liquor Store	2,000	-	97,000	12,000	3801 Harrodsburg Rd
Liquor Store	2,000	-	97,000	12,000	3765 Palomar Centre Dr
Steakhouse	1,123	1	6,800	46,000	455 Southland Dr
Steakhouse	839	1	24,000	46,000	3191 Beaumont Centre Cir
Steakhouse	2,000	-	320	46,000	101 Virginia Avenue

5.2. Final location-category pairs visualization

Shown in Figure 7 are the final chosen locations for location-category pairs again shown overlaid with the heatmap of existing visited Keenland venues. Most of the pairs are closely aligned with clusters of visited locations. However, the steakhouse venue on Virginia Avenue, shown as the only singular marker, does not seem to be aligned with visited areas, although it is not too far from several clusters both downtown and along Harrodsburg Road. This is reflected in Table 2, as this location has significantly fewer visits in a 1 km radius. Although there is no venue nearby, the lack of nearby visits may indicate that this pair is less favorable than other locations. As previously described, many of the other pairs are clustered in similar areas that align with clusters of visited venues.

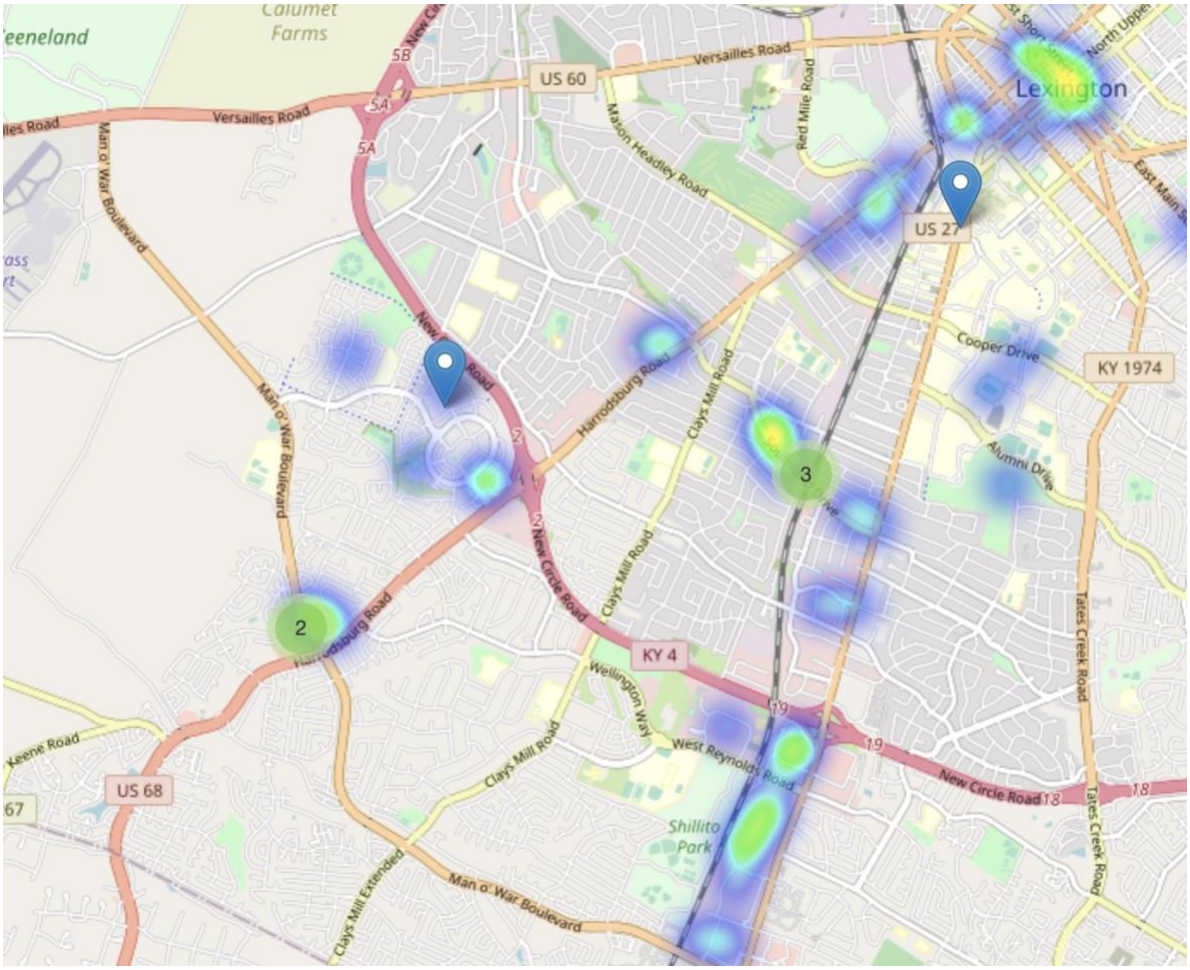

Figure 7: Final chosen location-category pairs overlaid on visited venue heatmap

Section 6. Discussion

The impact of Keeneland on Lexington's economy is vast and affects numerous areas based on the analysis of next venues. However, the number of visits is highly concentrated in a few areas, which impacts the relative favorability of location choice in developing a new venue. Additionally, the number of visits is heavily dominated by a few categories. These categories vary widely in their distance from studied pairs and number of venues within 1 km of the studied pairs. However, the majority of most ideal locations did not have any nearby venues in their categories.

While there are relatively few locations category pairs that are not inherently inferior based on the metrics studies, these pairs all have varying strengths and weaknesses. The decision on which venue to develop will depend on how the developer weighs the importance of each metric and what type of venue the developer plans to open. If catering more to the Keeneland crowd specifically is deemed more important, the number of visits within 1 km and the number of visits within each category would be most relevant. Based on number of visits in the area, opening a liquor store at either of the two nearby locations would be preferable. Alternatively, if the developer is concerned primarily with opening a venue that is proven popular among visitors to Keeneland, opening a steakhouse would be most ideal.

While developers looking to attract visitors from Keeneland must consider popularity of the category and area among Keeneland visitors, they must also consider how heavily concentrated that location is with venues of the category already. Thus, if a developer is more concerned with how feasible a venue is based on the number of venues of that category nearby and not frequented by Keeneland goers, the number of venues in that category within 1 km and the minimum distance to a venue in that category would be more important. Based on these metrics, all venues except for the steakhouse on Southland Drive and on Beaumont Centre Circle would be ideal as no venues in that category are within 1 km.

Section 7. Conclusion

The focus of this project was determining an ideal location and category for placement of a venue catering to visitors from Keeneland. To determine this location, the number of visitors from Keeneland to various venues in Lexington was first determined using the Foursquare API next venues query tool and a weighing technique. The resulting weights and results were then grouped and clustered to determine ideal locations for venue placement. Locations were generated at 300 m intervals up to 1.5 km away from the cluster center.

The distance from each visited venue to each location was measured, and the categories of venues within 1 km were determined. In addition, the categories of visited venues were grouped

and ordered based on number of visitors. The top categories were chosen and the top categories of visited venue not located within 1 km of each location were found.

After the top categories using visited venues that were not located near each point were found, the study was broadened to include all venues within 1 km. The Foursquare API nearby venues query tool was used to find all such venues nearby. Then, these venues were grouped by category, and the minimum distance of each location to the category and the total number of venues of that category roughly within 1 km were calculated.

Based on the number of visits within 1 km, number of visits in the category, minimum distance to a venue in the category, and the number of venues of the category roughly within 1 km, the final location-category pairs were determined. These pairs were found by filtering out pairs that were inherently inferior to other pairs using these four metrics. The pairs are shown in Table 2. These pairs each have varying strengths and weaknesses that must be weighed by the developer to determine final placement for a new venue. The focus of the venue can be used as the basis for deciding a final venue location.

In addition to these metrics, additional metrics not included in this study can be used as the basis for deciding upon a final location. These metrics include the demographics of the studied neighborhood, the number of visits in the area for all visitors, not just those from Keeneland, and the availability of each location based on which businesses are already there and the willingness of the business owner to sell. However, this study focused solely on the impact from Keeneland, and as such provided developers with a number of venues with which to capitalize on the extensive impact of the racetrack on the local economy.

Section 8. Bibliography

Bollinger, C. R. (2015). *A Measure of the Economic Impact of Keeneland Racing and Sales on Lexington-Fayette County.* Lexington: University of Kentucky.

Breeders' Cup. (2015, November 17). *Press Releases*. Retrieved from breederscup.com: https://www.breederscup.com/media-center/press-releases/2015-11-17

Foursquare. (2020). *Documentation, Places API*. Retrieved from developer.foursquare.com: https://developer.foursquare.com/docs/places-api/

Google. (2020, June 5). *Geocoding API*. Retrieved from developers.google.com: https://developers.google.com/maps/documentation/geocoding/start?utm_source=google&utm_medium=cpc&utm_campaign=FY18-Q2-global-demandgen-paidsearchonnetworkhouseads-cs-maps_contactsal_saf&utm_content=text-ad-none-none-DEV_c-CRE_433476780160-ADGP_Hybrid+%7C

Keeneland. (2019, April 26). *Keeneland Closes 2019 Spring Meet With Strong Attendance And Wagering Results*. Retrieved from keeneland.com: https://www.keeneland.com/media/news/keeneland-closes-2019-spring-meet-strong-attendance-and-wagering-results

Keeneland. (2019, October 26). *Strong Attendance And Community Events Highlight Keeneland Fall Meet*. Retrieved from keeneland.com: https://keeneland.com/media/news/strong-attendance-and-community-events-highlight-keeneland-fall-meet

LFUCG. (2020, January 29). *Zoning*. Retrieved from data.lexingtonky.gov: https://data.lexingtonky.gov/dataset/zoning

LFUCG Planning Office. (2020). *Zoning Ordinance Article 8: Schedule of Zones.* Lexington, KY: LFUCG.

Staff, H. L. (2019, June 26). Record purses, 10 Breeders' Cup berths await Keeneland Fall Meet competitors. *Lexington Herald Leader*.

US Census Bureau. (2019, July 1). *QuickFacts Lexington-Fayette, Kentucky*. Retrieved from census.gov: https://www.census.gov/quickfacts/lexingtonfayettekentucky

US Census Bureau. (2019, June 20). *TIGER/Line Shapefile, 2017, state, Kentucky, Current Block Group State-based*. Retrieved from Data.gov: https://catalog.data.gov/dataset/tiger-line-shapefile-2017-state-kentucky-current-block-group-state-based