Food For Thought

SURVMETH 727 Term Project

Rebecca Oh, Alexa Zmich

December 5, 2019

Contents

Introduction . . . 1
Data . . . 1
Results . . . 4
Discussion . . . 11
Link to GitHub Repository . . . 12
References . . . 12

Introduction

Rising obesity rates in the United States have become a large concern, as obesity is now the leading preventable cause of death worldwide and is associated with many other causes of death. The dietary choices we make every day affect the potential of developing obesity, including the types of food establishments we choose to eat at. Thus, we are interested in seeing the possible effect these food establishments can have on our health, leading to our research question: "Do the main cuisines of food establishments affect obesity rates within Michigan?" In our study, we additionally seek to delve further into this question by analyzing geographic location (by county), which may also be impacting the obesity rates.

Prior research for obesity is heavily focused on fast food restaurants specifically. Studies have found that the location of fast food reference, in regards to both quantity and closeness to heavily frequented buildings (i.e. schools), can have a significant effect on the obesity rates in the area (Jeffery et al. 2006; Davis and Carpenter 2011). The significance found on location greatly influenced our decision to include a location focus along with our research question, to see if results would be different for non-fast-food restaurants. Currie et al. (2010) separated fast food restaurants from non-fast-food restaurants and found there to be a lack of correlation between non-fast-food restaurants and obesity, however, the study focused primarily on the fast food restaurants and did not further break down the non-fast-food-restaurants due to the initial lack of correlation. Our research question will continue research into a much needed focus on non-fast-food-restaurants, especially as Yelp tends to lack full inclusion of area fast food restaurants.

Data

Cuisines and Counties Data

In order to answer our main research question, we first needed to gather data regarding main cuisines in Michigan counties. We chose Yelp as our data gathering source, as it stores key restaurant information beyond the restaurant name, such as cuisine type and location, which will be used to test for significant effects against obesity. After gaining access to Yelp API, we stored the key as an object to be used in the following API queries that ultimately formed our data set.

```
#create a token for use with API request
client_id <- "tuUnFmtQQZAX7YHwwbdz6g"
client_secret <- "PQL5AXms_2Yr6szyGBy2V6yH1sdeFS1g5S3od1WZxFKAD3IK0DEHC0mt1B8atTzO5mspVGcqxvA9gbAb_3u3_u07u

res <- POST("https://api.yelp.com/oauth2/token",
            body = list(grant_type = "client_credentials",
                        client_id = client_id,
                        client_secret = client_secret))

token <- content(res)$access_token
```

We encountered a few unexpected challenges with using the Yelp API. Gathering data by searching a whole county within the US is not supported, insisting that users search for a specific city. Additionally, the API has a limit of 50 queries. Thus, we collected the top 20 most popular food establishments within (or near) the most populated city for each of the 83 counties in the state of Michigan. We searched "food" as the term of interest and the location that we searched was the most populated city for each county. Then, we appended all 83 datasets to have a full dataset listing the most popular food establishments, the cuisine type of each establishment, and the city each establishment is located in.

The data from Yelp provide us information regarding location (city, state, longitude, latitude), but lacks county. Since we were ultimately hoping to analyze the data at the county level, we needed to create a new object providing us with this information, by mapping from the establishment's corresponding city. We will show an example of this using Wayne County:

```
#search url and collect data per county
yelp <- "https://api.yelp.com/v3/businesses/search/?limit=20&offset=0&sort=0&term=food&location=detroit+MI"
url <- modify_url(yelp, path = c("v3", "businesses", "search"))
res <- GET(url, add_headers('Authorization' = paste("bearer", client_secret)))
results <- content(res)

#parse data
yelp_httr_parse <- function(x) {

  parse_list <- list(id = x$id,
                     name = x$name,
                     categories = x$categories,
                     rating = x$rating,
                     review_count = x$review_count,
                     latitude = x$coordinates$latitude,
                     longitude = x$coordinates$longitude,
                     address1 = x$location$address1,
                     city = x$location$city,
                     state = x$location$state,
                     distance = x$distance)

  parse_list <- lapply(parse_list, FUN = function(x) ifelse(is.null(x), "", x))

  df <- tibble(id=parse_list$id,
                   name=parse_list$name,
                   categories = parse_list$categories,
                   rating = parse_list$rating,
                   review_count = parse_list$review_count,
                   latitude=parse_list$latitude,
                   longitude = parse_list$longitude,
                   address1 = parse_list$address1,
                   city = parse_list$city,
                   state = parse_list$state,
```

```
                    distance= parse_list$distance)
  df
}

results_list <- lapply(results$businesses, FUN = yelp_httr_parse)
food <- do.call("rbind", results_list)
county <- rep("Wayne",length(food$city))
food_data <- cbind(food, county)
```

We then repeated this process for the remaining 82 counties in Michigan.

To have a complete dataset, we appended as follows, and called our final merged dataset for cuisines and counties: "all _establishments."

```
#append 83 datasets
all_establishments <- rbind(
  food_data, food_data2, food_data3, food_data4, food_data5, food_data6,
  food_data7, food_data8, food_data9, food_data10, food_data11, food_data12,
  food_data13, food_data14, food_data15, food_data16, food_data17, food_data18,
  food_data19, food_data20, food_data21, food_data22, food_data23, food_data24,
  food_data25, food_data26, food_data27, food_data28, food_data29, food_data30,
  food_data31, food_data32, food_data33, food_data34, food_data35, food_data36,
  food_data37, food_data38, food_data39, food_data40, food_data41, food_data42,
  food_data43, food_data44, food_data45, food_data46, food_data47, food_data48,
  food_data49, food_data50, food_data51, food_data52, food_data53, food_data54,
  food_data55, food_data56, food_data57, food_data58, food_data59, food_data60,
  food_data61, food_data62, food_data63, food_data64, food_data65, food_data66,
  food_data67, food_data68, food_data69, food_data70, food_data71, food_data72,
  food_data73, food_data74, food_data75, food_data76, food_data77, food_data78,
  food_data79, food_data80, food_data81, food_data82, food_data83)
```

One of the columns of interest in "all_establishments," 'categories,' requires some formatting clean-up. The 'categories' used by Yelp were very specific, whereas we were hoping to use the more general cuisine types for our analyses. Additionally, we needed to place the categories into new, appropriate buckets to make the analyses relevant. Thus, we reformatted that column in order to make it more understandable and clear. We show a few examples of classifying 'categories' in the following chunk:

```
#clean up 'categories'
all_establishments$categories <- all_establishments$categories %>%
  str_replace_all(., "list\\(alias = ", "") %>%
  str_replace_all(., "\\)", "") %>%
  str_replace_all(., ", title = ", "") %>%
  str_replace_all(., "\"", "") %>%
  str_replace_all(., "burgersBurgers", "American") %>%
  str_replace_all(., "newamericanAmerican \\(New", "American") %>%
  str_replace_all(., "beergardensBeer Gardens", "Bars/Breweries")
```

Our initially cleaned data set of cuisines and counties (all_establisments) had 1568 observations, since the most populated areas of certain counties had less than 20 food establishment options on Yelp. However, when looking through this dataset, we noticed more areas of data-cleaning that needed to be addressed. For instance, Yelp has captured establishments that don't align with our restaurant-specific analyses (i.e. bowling). We grouped these establishments into a category called "Non-restaurant." We pre-determined to not include grocery stores ("Grocery" in our dataset) in our analyses as they are a fairly different food establishment type than a restaurant and it can be more difficult to determine the choices a consumer made or the impact those choices will have on their obesity. Lastly, we had a small group of establishments that did not fit into any categories, so we placed them in a category we called "Other." As these three categories may not be very useful in our analyses, we removed them from our dataset, shown below:

```
all_establishments <- all_establishments %>%
  filter(!categories == "Non-restaurant") %>%
  filter(!categories == "Grocery") %>%
  filter(!categories == "Other")
```

After removing the "Non-restaurant", "Grocery", and "Other" categories from all_establishments, our final dataset of cuisines and counties includes 1479 unique observations.

Obesity Data

Obesity data was collected from the Institute for Health Metrics and Evaluation website, in the form of an Excel file (saved as a CSV) listing counties with both a female and male obesity rate for each county. The file was then uploaded to our Github repository and we read in the raw url file.

```
obesity <- read_csv("https://raw.githubusercontent.com/daraeoh/survmeth727/master/obesity.csv")
```

We took the average of the male and female obesity rates and created a new column called "obesity_avg" to use as our depended variable of interest, as shown below. We will refer to this variable as "aggregate rate of obesity."

```
obesity_avg <-
  (obesity$female_obesity_prevalence_percent + obesity$male_obesity_prevalence_percent) / 2
avg_obesity <- cbind(obesity, obesity_avg)
```

Master Data

Now that we have gathered both the "cuisine and counties" data set and the "obesity" data set, we merged them together by "county" to get our master dataset (entitled "master") for analysis.

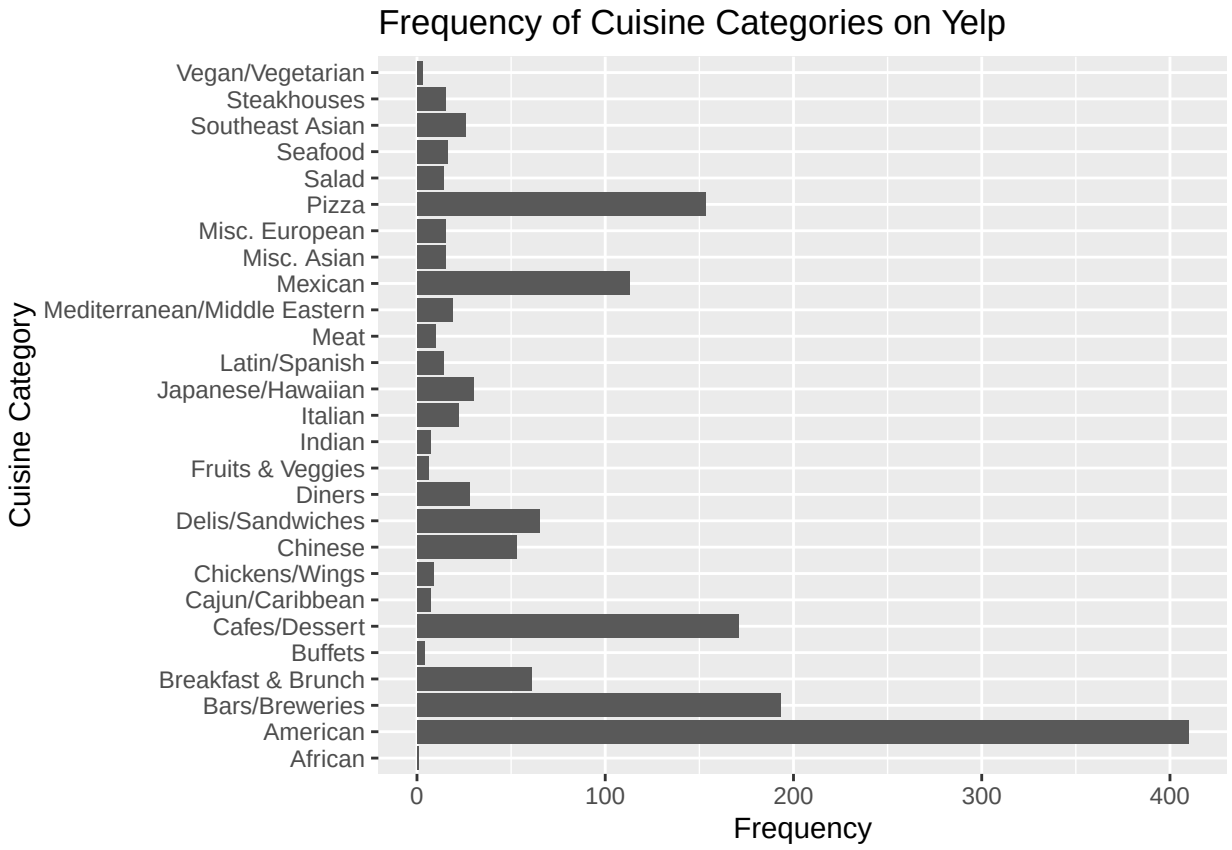
```
master <-inner_join(all_establishments, avg_obesity, by = "county")
```

Results

Data Exploration

In our exploratory analyses, we started by creating a basic barplot for "categories" to get an idea of the prevalence of different types of the most popular cuisines on Yelp for the state of Michigan.

```
#basic barplot
ggplot(master) +
  geom_bar(aes(x = categories)) +
  labs(
    x = "Cuisine Category",
    y = "Frequency",
    title = "Frequency of Cuisine Categories on Yelp"
    ) +
  coord_flip()
```



According to our barplot; American, Bars/Breweries, Cafes/Dessert, and Pizza are the most popular types of cuisine on Yelp for Michigan.

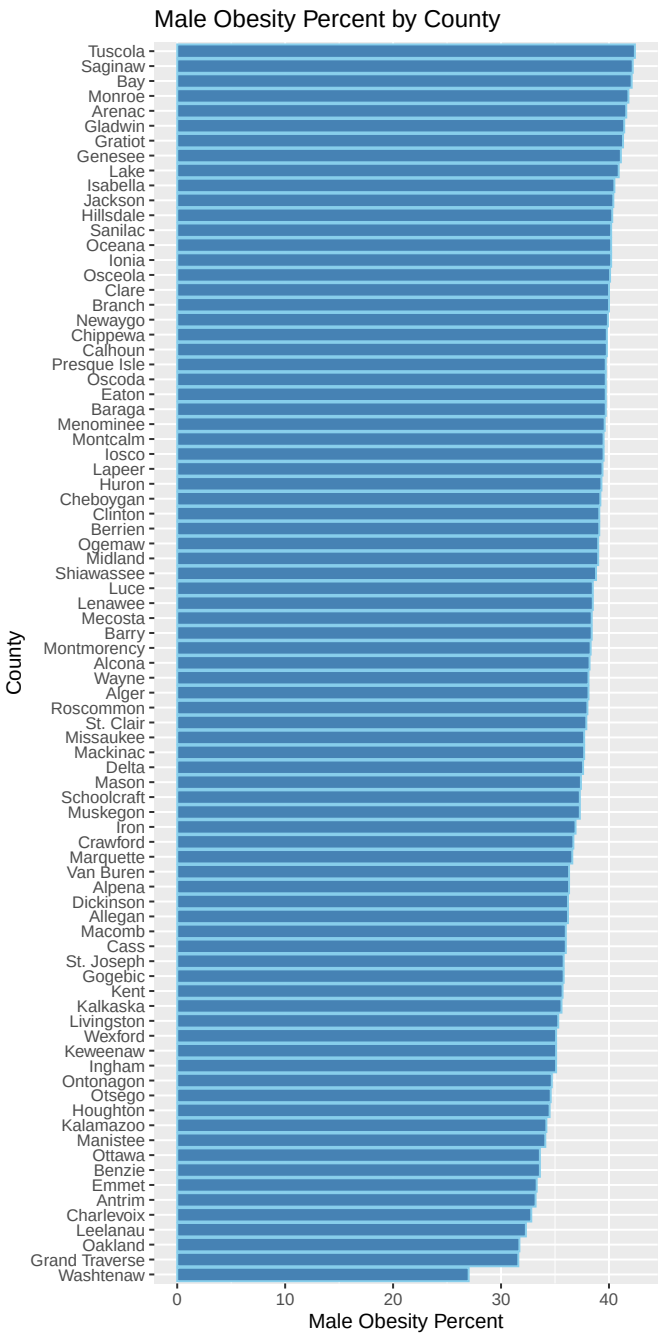
Since the obesity data we collected outlined the obesity rates for both males and females, we were also interested in seeing how the separate sexes obesity rates differed by county. We created bar plots outlining the sex obesity percent rate by county. The bar graphs listed below show that each sex had a fairly different line-up for obesity percent by county.

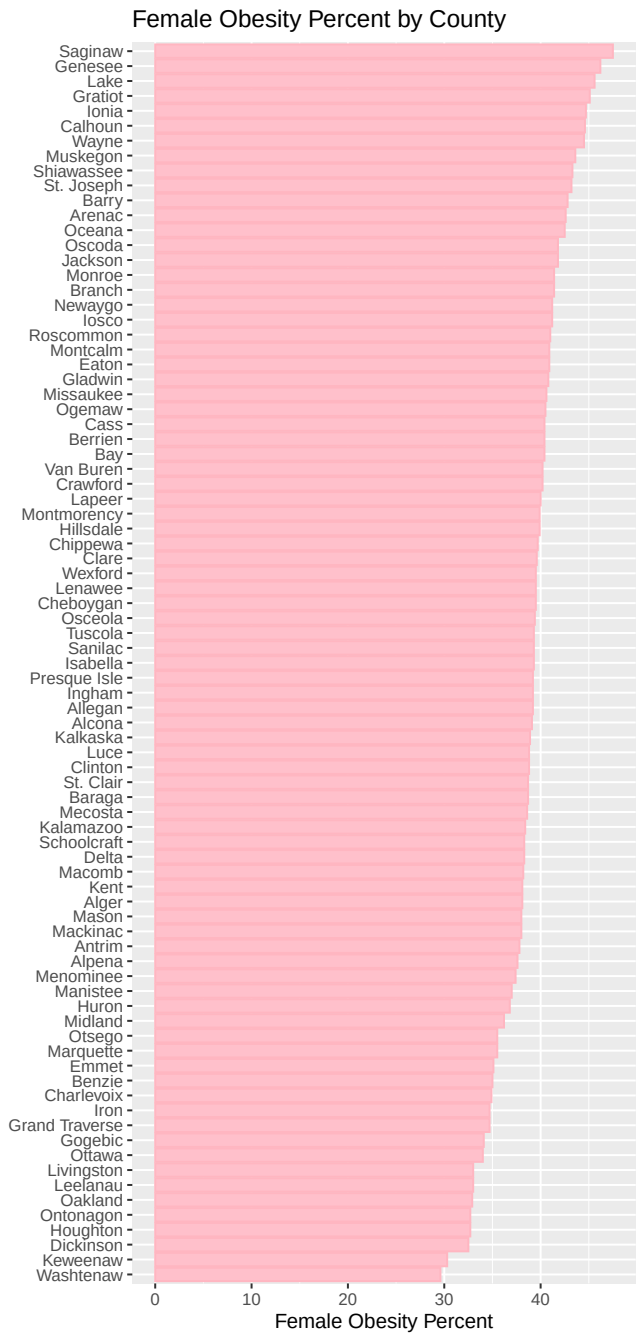
```
#male obesity barplot by county
male <- aggregate(
  male_obesity_prevalence_percent ~ county, master, mean) %>%
  arrange(desc(male_obesity_prevalence_percent))
male_plot <- ggplot(male,
                    aes(x = reorder(county, male_obesity_prevalence_percent),
                        y = male_obesity_prevalence_percent)) +
  geom_bar(
    stat = "identity",
    color = 'skyblue',
    fill = 'steelblue') +
  labs(
    x = "County",
    y = "Male Obesity Percent",
    title = "Male Obesity Percent by County"
    ) +
  coord_flip()

#female obesity barplot by county
female <- aggregate(
  female_obesity_prevalence_percent ~ county, master, mean) %>%
  arrange(desc(female_obesity_prevalence_percent))
```

```r
female_plot <- ggplot(female,
                      aes(x = reorder(county, female_obesity_prevalence_percent),
                          y = female_obesity_prevalence_percent)) +
  geom_bar(
    stat = "identity",
    color = 'lightpink',
    fill = 'pink') +
  labs(
    x = "",
    y = "Female Obesity Percent",
    title = "Female Obesity Percent by County"
    ) +
  coord_flip()

#print male and female obesity barplots
grid.arrange(male_plot, female_plot, ncol = 2)
```





Next, we were interested in observing the top cuisine categories with the highest values of average aggregate rate of obesity. In order to obtain these statistics, we first found the mean percentage of aggregate rate of obesity for each cuisine category. Then we listed the top ten cuisine categories with the largest percentage of aggregate rate of obesity for each cuisine category. When we listed the top ten cuisine categories with the largest percentage of aggregate rate of obesity, we found that "African," "Fruits and & Veggies," "Diners," "Meat," and "Indian" cuisines were among the cuisine categories that yielded the highest obesity rates, as shown in the table below. It is worth noting that "African" and "Fruits & Veggies" were both cuisine types with a small sample size, which could be affecting the received results.

```
#top 10 cuisine categories arranged by highest average aggregate rate of obesity
top_categories <- aggregate(
  obesity_avg ~ categories, master, mean) %>%
  arrange(desc(obesity_avg))
head(top_categories, 10)
```

Cuisine Categories	Average Aggregate Rate of Obesity (%)
African	40.87500
Fruits & Veggies	39.97500
Diners	39.41268
Meat	39.00550
Indian	38.99000
Buffets	38.98750
Chinese	38.93091
Pizza	38.71260
Delis/Sandwiches	38.64617
Bars/Breweries	38.47065

Similarly, we were interested in observing the top counties with the highest values of average aggregate rate of obesity, so we first found the mean percentage of aggregate rate of obesity for each county and then listed the top ten counties with the largest percentage of aggregate rate of obesity. These counties are listed in the table below:

```
#top 10 counties arranged by highest average aggregate rate of obesity
top_counties <- aggregate(
  obesity_avg ~ county, master, mean) %>%
  arrange(desc(obesity_avg))
head(top_counties, 10)
```

County	Average Aggregate Rate of Obesity (%)
Saginaw	44.85
Genesee	43.65
Lake	43.25
Gratiot	43.20
Ionia	42.45
Calhoun	42.20
Arenac	42.10
Monroe	41.60
Oceana	41.35
Wayne	41.30

Analysis

Based on our data exploration and our research question, we decided to fit linear models to our data to analyze the relationship of the variables we collected on the obesity rates. In order to find the model that best fit our data, we began with a simple approach, modeling the obesity average as our dependent variable against the single "categories" variable.

```
#model with 'cateogories' as the only predictor
master %>%
  lm(obesity_avg ~ factor(categories),
     data = .) %>%
  tidy()
```

```
#code including R-squared value
model1 <- lm(obesity_avg ~ factor(categories),
             data = master)
summary(model1)
```

term	estimate	p.value
(Intercept)	40.875000	1.654985e-74
factor(categories)American	-2.599681	2.193781e-01
factor(categories)Bars/Breweries	-2.404350	2.574884e-01
factor(categories)Breakfast & Brunch	-3.141583	1.433369e-01
factor(categories)Buffets	-1.887500	4.653939e-01
factor(categories)Cafes/Dessert	-2.473144	2.442854e-01
factor(categories)Cajun/Caribbean	-2.819444	2.271220e-01
factor(categories)Chickens/Wings	-3.175000	1.698905e-01
factor(categories)Chinese	-1.944091	3.657209e-01
factor(categories)Delis/Sandwiches	-2.228828	2.985644e-01
... with 17 more rows		

The R-squared value of our preliminary model was quite low (R-squared = 0.048), indicating poor model fit. To improve upon the model, we added county to the same model as above.

```
#model with 'categories' and 'county' as the predictors
master %>%
  lm(obesity_avg ~ factor(categories) + factor(county), data = .) %>%
  tidy()
```

```
#code including R-squared value
model2 <- lm(obesity_avg ~ factor(categories) + factor(county), data = master)
summary(model2)
```

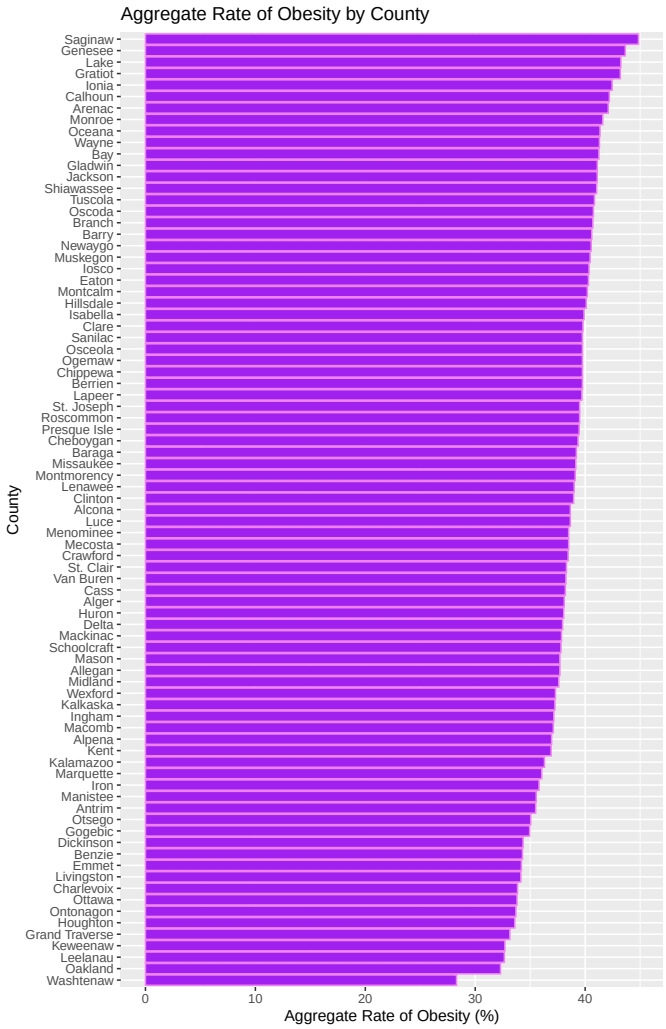
term	estimate	p.value
(Intercept)	3.865000e+01	0.000000000
factor(categories)American	-1.073765e-13	0.021576634
factor(categories)Bars/Breweries	-1.340980e-13	0.004236776
factor(categories)Breakfast & Brunch	-1.164266e-13	0.014129448
factor(categories)Buffets	-4.075038e-14	0.475383464
factor(categories)Cafes/Dessert	-1.227356e-13	0.009033133
factor(categories)Cajun/Caribbean	-8.744369e-14	0.090267795
factor(categories)Chickens/Wings	-1.355913e-13	0.008058154
factor(categories)Chinese	-9.941365e-14	0.036456553
factor(categories)Delis/Sandwiches	-1.062960e-13	0.024849817
... with 17 more rows of cuisine categories		
factor(county)Alger	-5.500000e-01	0.000000000
factor(county)Allegan	-9.500000e-01	0.000000000
factor(county)Alpena	-1.700000e+00	0.000000000
factor(county)Antrim	-3.150000e+00	0.000000000
factor(county)Arenac	3.450000e+00	0.000000000
factor(county)Baraga	5.500000e-01	0.000000000
factor(county)Barry	1.950000e+00	0.000000000
factor(county)Bay	2.600000e+00	0.000000000
factor(county)Benzie	-4.350000e+00	0.000000000
factor(county)Berrien	1.100000e+00	0.000000000
... with 73 more rows of counties		

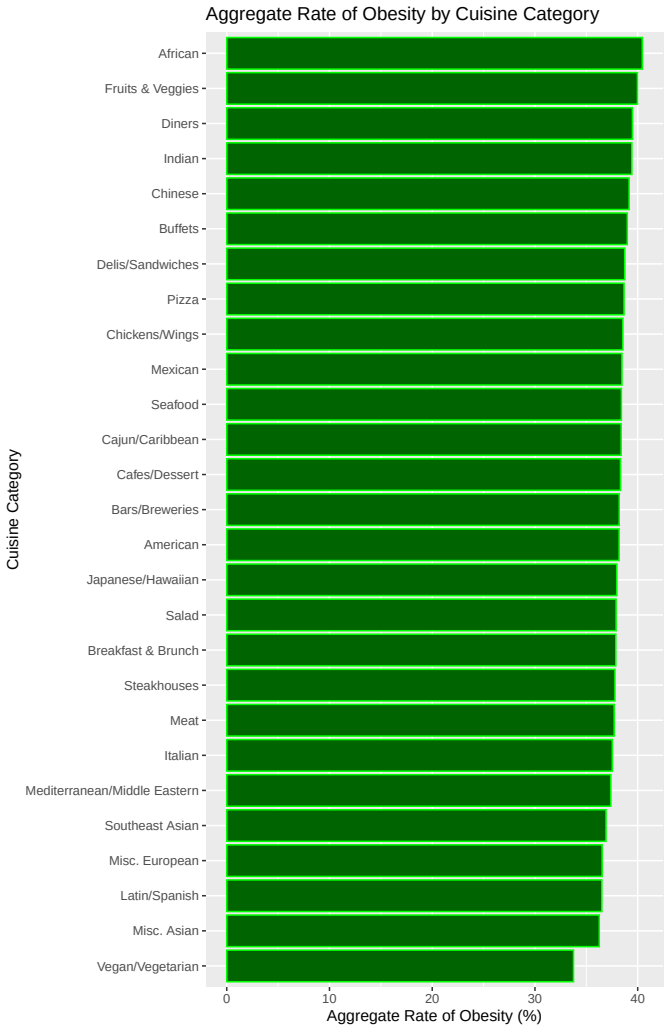
The R-squared value of the second model was much higher than the first (R-squared = 1), indicating a better model fit. Due to this, we decided to make two additional bar plots outlining the aggregate obesity rate by county and cuisine, respectively. We found the bar plots to be an interesting comparison to the results featured in our models.

```r
#aggregate rate of obesity plot by county
countyobes <- aggregate(
  obesity_avg ~ county, master, mean) %>%
  arrange(desc(obesity_avg))
county_plot <- ggplot(countyobes,
                      aes(x = reorder(county, obesity_avg),
                          y = obesity_avg)) +
  geom_bar(
    stat = "identity",
    color = 'violet',
    fill = 'purple') +
  labs(
    x = "County",
    y = "Aggregate Rate of Obesity (%)",
    title = "Aggregate Rate of Obesity by County"
    ) +
  coord_flip()

#aggregate rate of obesity plot by cuisine category
catobes <- aggregate(
  obesity_avg ~ categories, master, mean) %>%
  arrange(desc(obesity_avg))
categories_plot <- ggplot(catobes,
                      aes(x = reorder(categories, obesity_avg),
                          y = obesity_avg)) +
  geom_bar(
    stat = "identity",
    color = 'green',
    fill = 'darkgreen') +
  labs(
    x = "Cuisine Category",
    y = "Aggregate Rate of Obesity (%)",
    title = "Aggregate Rate of Obesity by Cuisine Category"
    ) +
  coord_flip()

#print both plots
grid.arrange(county_plot, categories_plot, ncol = 2)
```





Discussion

Our first model explored the effect of cuisine on obesity, without including the county variable. This model had poor fit and the variable outputs implied a lack of significance for the effect of cuisine on obesity, with the exception of three cuisine types: "Misc. Asian," "Misc. European," and "Vegan/Vegetarian." Adding county data into the model helped to explain the data in a better manner, as we expected. At a 0.05 threshold for significance, the majority of the cuisine types had a significant effect on the model, with the exception of the following cuisine types: "Buffets," "Cajun/Caribbean," "Mediterranean/Middle Eastern," "Misc. Asian," "Steakhouses," and "Vegan/Vegetarian." All the counties in Michigan, except for Luce county, had a significant effect on obesity. Our results indicate evidence to support the main focus of our research question exploring the effect of cuisine types on obesity rates. Additionally, the strong significance of the county data on the obesity rates reveals potential focus for future data.

A few limitations are throughout our research. Working with the Yelp API proved to be difficult for our dataset in a few different ways. As Yelp is constantly changing due to user generated input, our results would slightly change as well whenever we would re-run a model. Additionally, the cuisine data was not as concise as we had originally expected, which did lead to small sample sizes for each cuisine type. Though our cuisine types were not significant, future research with much larger sample sizes would still benefit from testing cuisine types to see if different results are received with more available data. Lastly, our model does not include weights, which does not account for the vastly different population sizes of some of the counties and the lack of available restaurants in the smaller counties.

In addition to the recommendations listed above, future research could benefit from exploring the effect of sex on obesity. Our data exploration found women to generally have higher obesity rates than men. When comparing the "Obesity Percent by County" bar plots for each sex to the "Aggregate Rate of Obesity by County," the "Aggregate Rate of Obesity by County" bar plot is closer aligned with the female plot than the male plot. Our research

question specifically focused on restaurant and location level data, leading to our decision to not include health and socio-demographic data, as that would alter the focus of our research.

Link to GitHub Repository

https://github.com/daraeoh/survmeth727

References

Currie, Janet, Stefano DellaVigna, Enrico Moretti, and Vikram Pathania. 2010. "Effect of Fast Food Restaurants on Obesity and Weight Gain."

Davis, Brennan, and Christopher Carpenter. 2011. "Proximity of Fast-Food Restaurants to Schools and Adolescent Obesity."

Institute for Health Metrics and Evaluations. (n.d.). US County Profiles. Retrieved from http://www.healthdata.org/us-county-profiles.

Jeffery, Robert W., Judy Baxter, Maureen McGuire, and Jennifer Linde. 2006. "Are Fast Food Restaurants an Environmental Risk Factor for Obesity?"