**RESEARCHING MIAMI-DADE COUNTY**

The purpose of this research is to determine the wealthy/poorer and tourist/non-tourist areas of Miami-Dade County. Then to relate that information to the data set. I believe this can be achieved by doing a general survey, then using that information to relate in to district, tract, and cluster of the data set.

**Key Questions:**

Where are the wealthy areas?

* What is considered to be wealthy?
  + Should this be tiered?

What are the Tourist areas?

* Will these be where the main industries are hotels?
* What about the cruise terminals? (I’m sure this affects hotels, car services in the area)
* What are the main attractions and where are they located.

**Helpful Websites:**

[**http://m-dcc.org/**](http://m-dcc.org/)

[**http://www.visitflorida.com/en-us/cities/miami.html**](http://www.visitflorida.com/en-us/cities/miami.html)

[**http://www.miamiherald.com/news/business/article182166466.html**](http://www.miamiherald.com/news/business/article182166466.html)

[**https://www.primebuyersreport.org/fl/miami-dade-county-fl-travel-agencies.html**](https://www.primebuyersreport.org/fl/miami-dade-county-fl-travel-agencies.html)

[**https://www.tripadvisor.com/Attractions-g34438-Activities-Miami\_Florida.html**](https://www.tripadvisor.com/Attractions-g34438-Activities-Miami_Florida.html)

[**http://www.touropia.com/tourist-attractions-in-miami/**](http://www.touropia.com/tourist-attractions-in-miami/)

[**https://www.timeout.com/miami/things-to-do/best-miami-attractions**](https://www.timeout.com/miami/things-to-do/best-miami-attractions)

[**https://www.planetware.com/tourist-attractions-/miami-us-fl-miami.htm**](https://www.planetware.com/tourist-attractions-/miami-us-fl-miami.htm)

[**http://www.miamiandbeaches.com/things-to-do/attractions**](http://www.miamiandbeaches.com/things-to-do/attractions)

**http://miamitours.com/maps/**

**Travel Agencies:**

**\*** [**https://www.bigbustours.com/en/miami/miami-bus-tours/**](https://www.bigbustours.com/en/miami/miami-bus-tours/) **-** 1 800 336 8233

**\* Miami Aqua Tours -** [(305) 358-7600](https://www.google.com/search?ei=R-osW7XbF8ausAWPhISYBQ&q=tours%20miami&oq=tours+miami&gs_l=psy-ab.3..0i7i30k1l10.16947.18026.0.18835.6.6.0.0.0.0.143.666.0j5.5.0....0...1c.1.64.psy-ab..1.5.664...35i39k1j0i67k1j0i131i67k1j0i20i263k1.0.f4nYxWWLtLk&npsic=0&rflfq=1&rlha=0&rllag=25775568,-80162019,2284&tbm=lcl&rldimm=1370310438296926970&ved=0ahUKEwi43JrYoefbAhXDmq0KHe0ZC2QQvS4ItgEwCg&rldoc=1&tbs=lrf:!2m1!1e2!2m1!1e3!3sIAE,lf:1,lf_ui:1)

\* Miami Urban Adventurs- [(305) 570-3934](https://www.google.com/search?client=psy-ab&hl=en&tbm=lcl&ei=kTg2W_31N8K7tgX1s7HoCg&q=miami+tour+guide&oq=miami+tour+&gs_l=psy-ab.3.2.0l10.525492.525747.0.528805.2.2.0.0.0.0.280.280.2-1.1.0....0...1c.1.64.psy-ab..1.1.278....0.SBg-dZ8ILKk)

\* City Sight Seeing Miami- [(305) 420-5538](https://www.google.com/search?client=psy-ab&hl=en&tbm=lcl&ei=kTg2W_31N8K7tgX1s7HoCg&q=miami+tour+guide&oq=miami+tour+&gs_l=psy-ab.3.2.0l10.525492.525747.0.528805.2.2.0.0.0.0.280.280.2-1.1.0....0...1c.1.64.psy-ab..1.1.278....0.SBg-dZ8ILKk)

Miami Tours- [(305) 444-4666](https://www.google.com/search?ei=fDg2W77xDJGAtQW697O4Cw&q=miami%20tours&oq=miami+tours&gs_l=psy-ab.3..0l8j0i10k1j0.1713.1923.0.3141.2.2.0.0.0.0.151.263.0j2.2.0....0...1c.1.64.psy-ab..0.2.263...0i7i30k1j0i8i7i30k1j0i30k1j0i8i10i30k1j0i8i30k1.0.VVHOJqhHpRs&client=psy-ab&hl=en&npsic=0&rflfq=1&rlha=0&rllag=25784386,-80135143,1246&tbm=lcl&rldimm=6543884868409099968&ved=0ahUKEwjmoem8gfnbAhVtCjQIHZCQBKUQvS4IWTAC&rldoc=1&tbs=lrf:!2m1!1e2!2m1!1e3!2m1!1e16!3sIAE,lf:1,lf_ui:1)

Miami Double Decker - [(305) 865-9999](https://www.google.com/search?ei=fDg2W77xDJGAtQW697O4Cw&q=miami%20tours&oq=miami+tours&gs_l=psy-ab.3..0l8j0i10k1j0.1713.1923.0.3141.2.2.0.0.0.0.151.263.0j2.2.0....0...1c.1.64.psy-ab..0.2.263...0i7i30k1j0i8i7i30k1j0i30k1j0i8i10i30k1j0i8i30k1.0.VVHOJqhHpRs&client=psy-ab&hl=en&npsic=0&rflfq=1&rlha=0&rllag=25784386,-80135143,1246&tbm=lcl&rldimm=6543884868409099968&ved=0ahUKEwjmoem8gfnbAhVtCjQIHZCQBKUQvS4IWTAC&rldoc=1&tbs=lrf:!2m1!1e2!2m1!1e3!2m1!1e16!3sIAE,lf:1,lf_ui:1)

**Miami Culinary Tours -** [(786) 942-8856](https://www.google.com/search?ei=fDg2W77xDJGAtQW697O4Cw&q=miami%20tours&oq=miami+tours&gs_l=psy-ab.3..0l8j0i10k1j0.1713.1923.0.3141.2.2.0.0.0.0.151.263.0j2.2.0....0...1c.1.64.psy-ab..0.2.263...0i7i30k1j0i8i7i30k1j0i30k1j0i8i10i30k1j0i8i30k1.0.VVHOJqhHpRs&client=psy-ab&hl=en&npsic=0&rflfq=1&rlha=0&rllag=25784386,-80135143,1246&tbm=lcl&rldimm=6543884868409099968&ved=0ahUKEwjmoem8gfnbAhVtCjQIHZCQBKUQvS4IWTAC&rldoc=1&tbs=lrf:!2m1!1e2!2m1!1e3!2m1!1e16!3sIAE,lf:1,lf_ui:1)

Tour Guide Miami- [(786) 487-7675](https://www.google.com/search?client=psy-ab&hl=en&tbm=lcl&ei=kTg2W_31N8K7tgX1s7HoCg&q=miami+tour+guide&oq=miami+tour+&gs_l=psy-ab.3.2.0l10.525492.525747.0.528805.2.2.0.0.0.0.280.280.2-1.1.0....0...1c.1.64.psy-ab..1.1.278....0.SBg-dZ8ILKk) north

<http://partners.miamiandbeaches.com/~/media/files/gmcvb/partners/research%20statistics/2017-gmcvb-annual-report>

Tourist Attractions:

* Lummus Park Beach
* Little Havana
* Little Haiti
* The beaches (South beach, midbeach, Miami beach)
* Bayside Market
* Art District
* Vizcaya Museum and Gardens
* Everglades National Park
* Calle Ocho
* Bayfront Park
* American Airlines Arena
* Zoo Miami
* Jungle Island
* Deering Estate
* Wynwood Walls Street Art
* Coral Castle
* Freedom Tower
* Miami Children's Museum
* Crandon Park
* Airports
* Venetian pools
* Venetian Way
* Miami Beach Botanical Garden
* Ocean Drive (street)
* Collins Ave (street)
* Lincoln Road
* Aventura Mall
* Downtown Miami
* Dolphin Mall
* Coconut Grove
* Key Biscayne
* Miami Seaquarium
* Dadeland Mall
* Miami International Mall
* Bal Harbour Shops
* Mall of Americas
* Auto Museum

We are looking at areas with beaches, shopping, and south beach/ ocean drive (pg 25). The areas below are where most visitors said they spent their time. The visitors also went to Art Deco District/ South Beach, Beaches, Lincoln Road, Bayside Market place, Downtown Miami, and Aventura Mall.

Match the most visited cities to their tracts, districts, and cluster. In the census these are called places. <http://blog.cubitplanning.com/2011/05/census-tracts-by-city/>

<http://www.usa.com/miami-dade-county-fl.htm>

<http://www.miamidade.gov/commission/districts.asp>

* Miami Beach (58.5%)
  + <https://www.miamibeachfl.gov/business/>
  + All tourist, many nature parks, a large church. Hotels line the beach, and many restaurants. An island.
  + Zip Codes:
    - 33140
    - 33139
    - 33141
  + Places:
    - Miami Beach is Cluster 2 (category) Mix between clusters 2 and 4 (issue type)
    - Venitian way is cluster 3 (cat) cluster 4 (isty)
    - South Beach is cluster 3 (cat) | Mostly cluster 4 (isty)
  + Wealthy/ Lower income areas:
    - Tract 3906:
      * Guessing tourist/ vacation home because of the house list price 1mill.
    - Tract 4102:
      * Appears as a little island, mostly White (93%)
    - Tract 3914:
      * Behind the Miami Beach strip
      * 30, 000 median income
      * Workers?
      * Small hotels but mostly chain restaurants and stores.
  + District 4 and 5
* City of Miami (Downtown Miami) (14.9%) (Little Havana, Little Haiti, etc), Coconut Grove (2%), Bayside Market Place.
  + Good portion in the bottom of District 3
  + Night life, Miami Heat games, Art district
    - <http://www.miamidade.gov/district03/places-of-interest.asp>
  + Business district
  + Zipcodes: 33133 (Coconut grove), 33129 (Museum/DTM/ bayside), 33132 (Cruise), 33131(DT), 33231(dt), 33130(calle 8), 33299(dt)
* North Dade/Sunny Isles Beach (8.1%)
  + Trump Towers
  + Relatively poor area
* Airport Area (10.9%)
* South Miami-Dade (2.4%)
* Doral (1.7%)
* Key Biscayne (1.5%)
* Coral Gables (1.9%)
* Aventura
  + 33180
  + The mall.

This file maps zip codes to census tracts!!!! Going from

<https://www.huduser.gov/portal/datasets/usps_crosswalk.html>

<http://www.verstaresearch.com/blog/merging-zip-codes-with-census-data/>

Demographic Statistics per tourist areas:

Mean Percents of each category

# tourist\_area per\_cap\_income under\_poverty unemployment\_rate white african\_american

#1 N.Tourist 24701.38 20.55778 8.736000 13.69867 18.40044

#2 Tourist 36951.94 15.12424 7.127273 27.93333 12.67576

#asian hispanic ed\_high college\_ed att\_college\_degree

#1 1.497778 65.54311 28.86844 17.17244 9.921333

#2 1.548485 44.30303 17.53939 21.15758 16.484848

Census\_Category csv- Mean of percent per demographic category.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Per Cap Income** | **Under Poverty** | **Unemploy. Rate** | **White** |
| **Tourist** | 36951.94 | 15.12424 | 7.127273 | 27.93333 |
| **N. Tourist** | 24701.38 | 20.55778 | 8.736000 | 13.69867 |
|  | **African American** | **Asian** | **Hispanic** | **High school** |
| **Tourist** | 12.67576 | 1.548485 | 44.30303 | 17.53939 |
| **N.Tourist** | 18.40044 | 1.497778 | 65.54311 | 28.86844 |
|  | **College** | **College Degree** |  |  |
| **Tourist** | 21.15758 | 16.484848 |  |  |
| **N. Tourist** | 17.17244 | 9.921333 |  |  |

Of the random sampled data 258/517 approximately 12.8% are tourist tract areas (Census Category csv).

Census\_Category

|  |  |  |
| --- | --- | --- |
| Cluster | Amount per cluster | Proportion |
| 1 | 4 | 0.121 |
| 2 | 8 | 0.242 |
| 3 | 9 | 0.273 |
| 4 | 12 | 0.364 |
|  |  |  |

* You saved a file (category311\_mean\_delay\_all\_tracts). This is from the sampled tracts.

The stuff below is the mean delay time for both tourist and non tourist tracts

# t n mean max min median s.dv

# 1 N.Tourist 286237 -19.4 1626 -15906 -4 74.1

# 2 Tourist 18844 - 6.53 924 - 365 -3 62.7

THIS IS ISSUE TYPE CSV

#tourist\_area per\_cap\_income under\_poverty unemployment\_rate white african\_american asian hispanic

#1 N.Tourist 24701.38 20.55778 8.736000 13.69867 18.40044 1.497778 65.54311

#2 Tourist 36951.94 15.12424 7.127273 27.93333 12.67576 1.548485 44.30303

#ed\_high college\_ed att\_college\_degree

#1 28.86844 17.17244 9.921333

#2 17.53939 21.15758 16.484848

cluster amount proportion

#<int> <int> <dbl>

# 1 1 7 0.212

# 2 2 3 0.0909

# 3 3 11 0.333

# 4 4 12 0.364

Top Calls per year cluster 4

Is there anything interesting about the cluster 4 311 calls. What are the characteristics of calls by time of year? Are the frequency of calls more or less in the summer months? Which calls are more frequent in which time of year? Have the calls per certain time of year re-occured over the past few years? How were calls received (maybe)?

2013 1184

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2013 | PET ACCOUNT UPDATE  (34.1%) | STRAY / DOG-AT-LARGE | INJURED ANIMAL | ANIMAL BITE TO A PERSON | DEAD ANIMAL PICKUP MD |
| 2014 | PET ACCOUNT UPDATE | STRAY / DOG-AT-LARGE | INJURED ANIMAL | ANIMAL BITE TO A PERSON | ANIMAL CRUELTY INVESTIGATION |
| 2015 | PET ACCOUNT UPDATE | ANIMAL BITE TO A PERSON | INJURED ANIMAL | STRAY / DOG-AT-LARGE | ANIMAL CRUELTY INVESTIGATION |
| 2016 | PET ACCOUNT UPDATE | INJURED ANIMAL | DEAD ANIMAL PICKUP MD | ANIMAL BITE TO A PERSON | STRAY / DOG-AT-LARGE |
| 2017 | PET ACCOUNT UPDATE | ANIMAL BITE TO A PERSON | ANIMAL CRUELTY INVESTIGATION | STRAY / DOG-AT-LARGE | INJURED ANIMAL |
| 2018\*\*\* | PET ACCOUNT UPDATE | DEAD ANIMAL PICKUP MD | INJURED ANIMAL | ANIMAL CRUELTY INVESTIGATION | ANIMAL SERVICES DOG TAG CHECK |

**Doing the t-test results:**

1. Resources- <https://www.r-bloggers.com/two-sample-students-t-test-1/>

* <https://suinotes.wordpress.com/2009/11/30/understanding-t-test-in-r/>
* <https://en.wikipedia.org/wiki/Welch%27s_t-test>
* <http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html>

1. Problems: Do we use the Welsh or regular two sample t test? Can we use both tests depending on the results of the variance test?

1. The Test results:
   1. For per capita income:

data: ntour\_data\_for\_t$per\_cap\_income and tour\_data\_for\_t$per\_cap\_income

t = -2.4447, df = 35.514, p-value = 0.01959

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-22418.242 -2082.881

sample estimates:

mean of x mean of y

24701.38 36951.94

* 1. For percent under poverty

data: ntour\_data\_for\_t$perc\_under\_poverty and tour\_data\_for\_t$perc\_under\_poverty

t = 2.0907, df = 37.92, p-value = 0.04331

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.1720416 10.6950291

sample estimates:

mean of x mean of y

20.55778 15.12424

* 1. For percent unemployment rate

data: ntour\_data\_for\_t$perc\_unemployment\_rate and tour\_data\_for\_t$perc\_unemployment\_rate

t = 1.2007, df = 36.356, p-value = 0.2376

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.107737 4.325192

sample estimates:

mean of x mean of y

8.736000 7.127273

* 1. Percent White

data: ntour\_data\_for\_t$perc\_white and tour\_data\_for\_t$perc\_white

t = -3.8094, df = 37.526, p-value = 0.0005013

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-21.802318 -6.667015

sample estimates:

mean of x mean of y

13.69867 27.93333

* 1. Percent African American \*

data: ntour\_data\_for\_t$perc\_aa and tour\_data\_for\_t$perc\_aa

t = 1.2983, df = 45.705, p-value = 0.2007

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-3.15276 14.60213

sample estimates:

mean of x mean of y

18.40044 12.67576

* 1. Percent Asian\*

data: ntour\_data\_for\_t$perc\_asian and tour\_data\_for\_t$perc\_asian

t = -0.13872, df = 256, p-value = 0.8898

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.7705253 0.6691111

sample estimates:

mean of x mean of y

1.497778 1.548485

* 1. Percent Hispanic\*

data: ntour\_data\_for\_t$perc\_hispanic and tour\_data\_for\_t$perc\_hispanic

t = 4.1333, df = 256, p-value = 4.85e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

11.12036 31.35981

sample estimates:

mean of x mean of y

65.54311 44.30303

* 1. Percent High school educated\*

data: ntour\_data\_for\_t$perc\_ed\_high and tour\_data\_for\_t$perc\_ed\_high

t = 5.2308, df = 38.122, p-value = 6.391e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

6.944984 15.713117

sample estimates:

mean of x mean of y

28.86844 17.53939

* 1. Percent College \*
     1. This made a difference when running the regular vs welch. Welch: p-value = 0.08934 and Reg: p-value = 0.02489

data: ntour\_data\_for\_t$perc\_ed\_college and tour\_data\_for\_t$perc\_ed\_college

t = -1.745, df = 36.804, p-value = 0.08934

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-8.6133871 0.6431244

sample estimates:

mean of x mean of y

17.17244 21.15758

* 1. Percent College degree

data: ntour\_data\_for\_t$perc\_ed\_att\_col\_deg and tour\_data\_for\_t$perc\_e\_att\_col\_deg

t = -2.8154, df = 37.117, p-value = 0.007753

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.286641 -1.840389

sample estimates:

mean of x mean of y

9.921333 16.484848

1. Data used is: sam\_issue\_census

**Rexamining the delayed variable**

1. Data set: mydata2

2.

Running the Regression again with the Citizenship variable

1. Date used: citizen\_data, perc\_citizen, cat2

Looking and Wealthy and Poverty Areas:

Data used: census\_call2

003706 0

2 007902 3.00

3 007807 3.60

4 010193 3.60

Higher

1 001803 63.1

2 002800 52.3

3 000807 50.2

4 000503 50.0

5 980300 50.0

\*\*\* Need to delete 003706 because there is no med or % under poverty…but there is 1368 calls from there.

Numbers per tract:

1803 2447

2 2800 946

3 7902 1226

4 10193 728

294337

Number of obs

July1\_new\_census\_data %>% select(TRACTCE10, tourist\_area, tot\_pop) %>% group\_by(tourist\_area) %>%

+ summarize(sum = sum(tot\_pop))

# A tibble: 2 x 2

tourist\_area sum

*<chr>* *<int>*

1 N.Tourist 1191693

2 Tourist 95966

# A tibble: 2 x 5

tourist\_area mean sd sum median

*<chr>* *<dbl>* *<dbl>* *<int>* *<int>*

1 N.Tourist -51.1 95.6 -2421037 -15

2 Tourist -50.4 104 - 169780 -12

July1\_new\_census\_data %>% select(tourist\_area, cluster) %>% group\_by(tourist\_area) %>% summarize(n = n())

# A tibble: 2 x 2

tourist\_area n

*<chr>* *<int>*

1 N.Tourist 225

2 Tourist 28

July1\_new\_census\_data %>% select(tourist\_area, cluster) %>%

+ filter(tourist\_area == "Tourist") %>% group\_by(cluster) %>% summarize(n = n())

# A tibble: 4 x 2

cluster n

*<int>* *<int>*

1 1 3

2 2 3

3 3 10

4 4 12

NON Cleaned data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | n | Percent of Delayed Calls | delayed time in days | Number of Calls delayed |
| N. Tourist | 277,923 | 17.055% | /number of calls delayed | 47,401 |
| Tourist | 18,225 | 20.636% | /number of calls delayed | 3,761 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | n | Percent of Delayed Calls | delayed time in days | Number of Calls delayed |
| N. Tourist | 277,923 | 17.055% | /number of calls delayed | 47,401 |
| Tourist | 18,225 | 20.636% | /number of calls delayed | 3,761 |

July 9th 2018:

The tasks are to: (1) understand dummy variables and how to use them in linear + logistic regression, (2) understand the output of logistic regression in terms of logs odd ratio and pseudo-r.

From the Political Analysis using R book:

* Page 101
* You have one outcome (1 or 0) and then predicter variables
* Have to use family=”binomial(link=”logit”)
* Logistic model allows us to approach in the simple linear model fashion but requires much more interpretation.
* AIC measures a good fit. It isn’t useful on its own but is good for comparing other models to each other, other predicter variables, or after removing one predicter variable.
  + Lower values have a better fit
* When creating a table do display. Put the N, the AIC and the % it predicted
* Odds ratio- the ratio of the probability the event occurs to the probability it does not occur.
  + Tells us the multiplicative factor by which the predicter variable will change on a one unit increase
  + Where the odds is an increase or decrease will depend on the estimate sign on the print out.
* Everything goes back to it being the outcome of 1

From this Youtube video to get me started <https://www.youtube.com/watch?v=fTfMdCQJz4s> :

* Dummy variables happen in categorical variables.
* Within these categorical variables there are -n- amount of categories.
* When rum in the regression the amount of categories present will always be n-1
* So I trandform the data into a wide-frame so that the observations become columns…?

From LIS 5802:

* The video uses one column and its numbers price @ 10, 20 and 30.
* However, the Miami 311 set uses 4 race columns and their percentages? How can we combine this? by a significant size population such as that particular race being above 50%? This would leave out significant proportions below that? How would we determine a threshold?
* What would be the proper way to combine the columns and still keep the percentages? Would we leave every race as a column and then categorize their percentage?
* Would be combine race into one column and just make a dummy variable out of it? How could R tell the percentages?
* How would I change the data formatting?

From youtube https://www.youtube.com/watch?v=2s8AwoKZ-UE :

* Make the white variable all equal to zero

July 12th

New measures for the mean percent with new variables

tourist per\_cap\_income perc\_unpov unemploy white black hispanic other highschool college coldegree other\_ed non\_cit

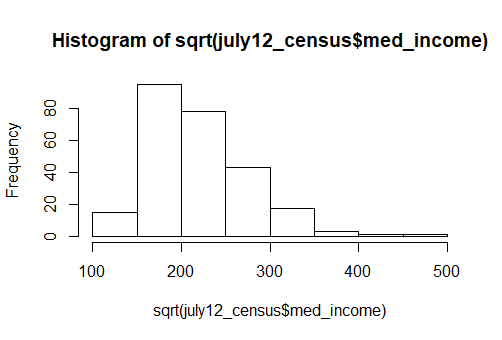
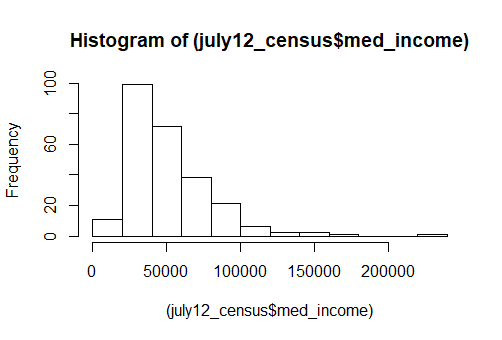
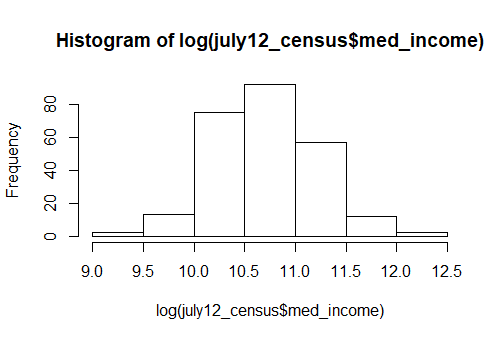
*<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 N.Tourist 24701 20.6 8.74 13.7 18.4 65.5 2.36 28.9 17.2 9.92 44.0 23.7

2 Tourist 43293 17.8 8.40 32.6 14.0 49.9 3.45 19.4 24.7 19.2 36.7 21.4

July 18: Transforming median income for regression model

* Use parametric statistics most of the time unless it is misleading. Parametric statistics allow us to see the differences between actual values opposed to ranks or ordinal positions.
* <http://www.unm.edu/~marcusj/datatransforms.pdf>
* In order to use parametric statistics data has to be on a consistent interval such as -2, -1, 0, 1, 2 however some data are not like this. A way to force the data like this or transform it is by logging or square rooting the data. This transforms it to another ratio scale where a true zero values exist.
* Can use either or/ the one that works on the distribution because bot
* These minimize the effect of outliers, where log function is more effective.
* Transforming works well with data that is skewed right (data concentrated on the left side of the graph)
* A-d test <https://rexplorations.wordpress.com/2015/08/11/normality-tests-in-r/>



* Median Income needs to be transformed before performing the regression model because its scale is much larger than the rest of the other input variables. This can be done through data normalization which takes skewed data and changes the scale ratio so the data are on a consistent interval [http://www.unm.edu/~marcusj/datatransforms.pdf]. There are several options for normalizing data such as logging and square rooting. While both can be used, logging has a greater effect on outliers and should be tried first [http://www.unm.edu/~marcusj/datatransforms.pdf]. The distribution displayed after creating a histogram using the raw data shows the median income variable (med\_income) is skewed right or has much of the data concentrated on the left tail of the graph. (*My first assumption is to remove the outliers and see if that has any impact on the data set however because the variable was med\_income I didn’t want to. This is because the outliers in this case -from what we know- are not mistakes in the data or due to faulty data entry. It also could be important when looking at if there are any differences between tourist and non tourist areas*). Transforming the data using the square root still results in a skewed graph. Lastly by logging the data, the histogram produced appears to be on a bell cure/ normal distribution.
* However, it is not enough to assume normality based on a histogram output. The Anderson-Darling test of normality will examine whether a sample is normally distributed [http://www.unm.edu/~marcusj/datatransforms.pdf]. A high p-value of 0.451 produced after running ad.test(med) on the logged values of med\_income shows there is a significant chance the data is normally distributed.
* #Anderson-Darling normality test

#data: med

#A = 0.35792, p-value = 0.451

July 18th Regression Model for number per tract:

#Call:

# lm(formula = number\_per\_tract.x ~ log\_med\_income + perc\_under\_poverty +

# perc\_unemployment\_rate + perc\_white + perc\_hispanic + perc\_ed\_college +

# perc\_non\_citizen + bi\_tourist)

#Residuals:

# Min 1Q Median 3Q Max

#-1436.5 -584.8 -221.1 342.8 4415.2

#Coefficients:

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) -3809.292 2898.022 -1.314 0.18993

#log\_med\_income 580.171 260.904 2.224 0.02709 \*

#perc\_under\_poverty 6.123 10.045 0.610 0.54272

#perc\_unemployment\_rate 26.079 19.053 1.369 0.17234

#perc\_white -16.666 6.681 -2.494 0.01328 \*

#perc\_hispanic 3.643 3.612 1.009 0.31414

#perc\_ed\_college -34.523 11.994 -2.878 0.00435 \*\*

#perc\_non\_citizen -39.248 8.214 -4.778 3.05e-06 \*\*\*

#bi\_tourist1 -148.108 222.970 -0.664 0.50716

#---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 1003 on 244 degrees of freedom

#Multiple R-squared: 0.2968, Adjusted R-squared: 0.2738

#F-statistic: 12.88 on 8 and 244 DF, p-value: 1.928e-15

* <https://data.library.virginia.edu/diagnostic-plots/>
* Pvalues and R ^2 show us how well the model represents the data
* <https://data.library.virginia.edu/understanding-q-q-plots/>
* Diagnostic tests of the model show that the model is non normally distributed or skewed. We can fix this by transforming more variables or switch models. Is the model bias?