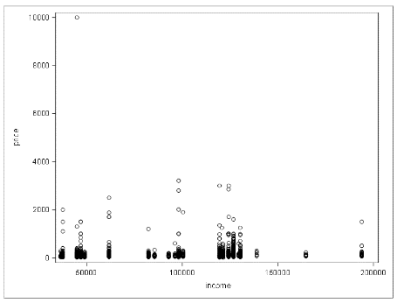
STA402: Final Project
Karen Gaither
5/11/2020

I chose to use the Airbnb data to see if the median income in each neighborhood of Manhattan has a significant effect on the price of the Airbnb listings in that area. I found data of the median income for the 50 most populous neighborhoods in Manhattan that I used to join with the Airbnb data set. I wanted to focus on Manhattan because that is the borough that has the most notable tourist attractions, such as Central Park, Times Square, and The Empire State Building. I suspect that a large amount of tourism could have a negative impact on housing for permanent residents of Manhattan. To investigate this I found data of the percentage of people on food stamps in each neighborhood. The income threshold to qualify for food stamps is just slightly above the poverty line so it is a good indication of the percentage of people in poverty, or close to it. I want to look, perhaps informally, at how the poverty rate is related to prices of Airbnb listings which could point to some gentrification.

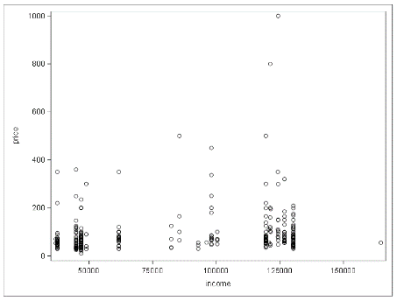
The data I found about income and food stamp rates were just presented on the website, so I had to manually put it into CSV format. This made it easy to read in and merge since it only included the information that I needed. The Airbnb dataset was also fairly straightforward to read in, and I dropped all the variables that I was not going to use to make it more manageable. The first challenge I faced was in merging these datasets. I first tried to use the merge statement in a data step but because there were multiple observations in the listings dataset and only one observation in the income data for each neighborhood it didn't work how I thought it would. I instead used proc sql to do an inner join so that only neighborhoods that were in both datasets were included in the final data. I did another inner join afterwards to merge the food stamps data with that as well.

While I wanted to focus on the relationship between income and price, I knew that the room type would probably have an effect on the price as well. An entire home being rented is most likely more expensive than a single room in the same area, so I had to account for these differences when conducting a test. I started by looking at the scatter plots of price vs. income for each room type. From the scatterplots alone there is not necessarily an obvious linear

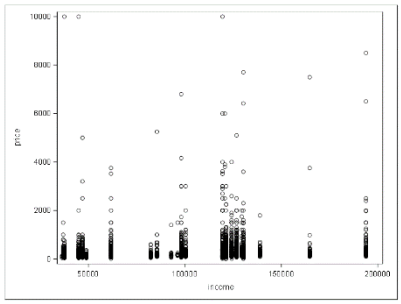
correlation between the two variables in any of the room types. There may be a positive relationship in the shared room and entire home categories, but again it is not very strong just from the plots.



Private Room

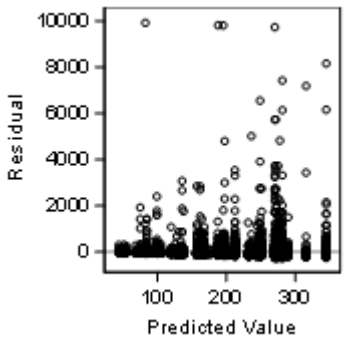


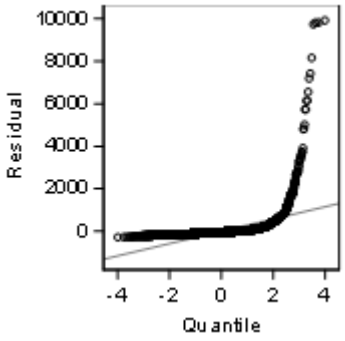
Shared Room



Entire Home/Apt

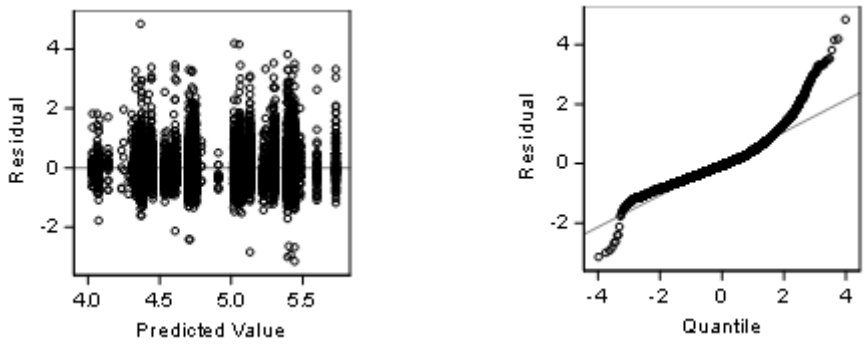
To test the hypothesis, I fitted a general linear model with price as a function of income and room type. While the results of this test indicated that both room type and income were significant, as the p-values were less than the chosen alpha level of 0.05 (Figure 1), both the constant variance and normality assumptions were violated so I could not use the model. In the residuals plot it is clear that the variances are increasing and the QQ plot is very curved meaning the data is not normal.





To account for these violations, I did a log transformation of the dependent variable. After adding the log of price to the data set, I fitted another general linear model using the transformed variable. Once again, this test indicated that room type and income are significant predictors of price because both p-values are less than the alpha value of 0.05 (Figure 2), and this time the

assumptions are met. There is no obvious pattern in the residuals plot and though there is some variance at the tails of the QQ plot there should not be any issues since it is such a large dataset.



Though room type and income are significant in predicting the price of an Airbnb, the interpretation of the results is more complicated because of the transformed variable. To determine how much the price increases as the income increases, I have to transform the model estimate value from the SAS output which is .000004537. Transforming this value by doing $(e^{.000004537} - 1) * 100$ gives the percent increase in price. For each $1 increase in median income, the price of an Airbnb in that Manhattan neighborhood is predicted to increase by 0.0004537%.

To look at the impact of poverty on prices I fitted a general linear model of price as a function of the food stamps rate and found that it is a significant predictor of price and has a negative correlation which is to be expected (Figure 3). It makes sense that neighborhoods with higher incomes and therefore higher Airbnb prices would have a lower percentage of people on food stamps, however, since the data is the median income and not the mean, it is not as indicative of outliers in either direction. This means that high Airbnb prices with a moderate income level and high poverty level could be an indication of gentrification in the area. In looking at a table with the average Airbnb price, median income, and percent of people on food stamps by neighborhood (Figure 4) there are a few points that stick out as unusual. For example, Little Italy has an average Airbnb price of $222.06 which puts it in the more expensive half of all neighborhoods, but 16.8% of residents are on food stamps and the median income is $85,500 which is much lower than other neighborhoods with comparable average prices. A similar pattern is seen in Chelsea, which is even more expensive than Little Italy. Though these few points may not make a large impact overall, there are signs that these neighborhoods are favoring middle class residents and tourists. A study done a few years ago found that for every 1%

increase in the number of Airbnb listings the price of rent in the area also increases by 0.018%. This relationship caters to tourists but depletes affordable housing options for residents of the city. There is not a definite test for if a neighborhood is undergoing gentrification or not since it is somewhat subjective in nature, but I think the relationship between income and food stamp rates seen here supports the idea that it is prevalent in Manhattan.

In conclusion, the price of Airbnb listings is significantly impacted by the median income within neighborhoods of Manhattan. As the income level increases the prices tend to increase as well; this is true for all room types: private room, shared room, and the entire home. The type of room is also significant, as a larger space would logically cost more money. Along with this, there are signs of gentrification in some of the neighborhoods due to a high price of the Airbnb coupled with a high food stamps rate. Areas that may have been more dominated by lower income residents in the past are shifting to please higher-class people which causes problems in the long run when it comes to making housing accessible for all.

Appendix:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
room_type	2	54481215.23	27240607.62	326.17	<.0001
income	1	25267224.10	25267224.10	302.54	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	6.4854798	B	16.00286017	0.41	0.6853
room_type Entire home/apt	144.1269153	B	15.58572881	9.25	<.0001
room_type Private room	30.9246489	B	15.70872406	1.97	0.0490
room_type Shared room	0.0000000	B	.	.	.
income	0.0010017		0.00005759	17.39	<.0001

1 - proc glm output; not transformed

Source	DF	Type III SS	Mean Square	F Value	Pr > F
room_type	2	2070.189171	1035.094585	3622.98	<.0001
income	1	518.368173	518.368173	1814.37	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	3.860477357	B	0.02959825	130.43	<.0001
room_type Entire home/apt	0.990617737	B	0.02882674	34.36	<.0001
room_type Private room	0.301553778	B	0.02905420	10.38	<.0001
room_type Shared room	0.000000000	B	.	.	.
income	0.000004537		0.00000011	42.60	<.0001

2 - proc glm output; log transformation

Source	DF	Type III SS	Mean Square	F Value	Pr > F
stamps_rate	1	38459770.10	38459770.10	441.25	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	254.3121750	3.57037284	71.23	<.0001
stamps_rate	-4.1418703	0.19717494	-21.01	<.0001

3 - proc glm output

Obs	neighborhood	avg_price	min_price	max_price	income	stamps_rate
1	Inwood	89.012	22	359	48900	35.0
2	Marble Hill	89.167	40	274	36300	40.4
3	Washington Heights	89.570	16	1000	44900	35.2
4	Roosevelt Island	113.260	30	1400	92900	11.7
5	Morningside Heights	114.904	30	1200	82300	8.6
6	Harlem	119.125	10	5000	46900	22.5
7	East Harlem	133.130	30	9999	37500	31.2
8	Chinatown	161.665	41	1500	36900	27.8
9	Stuyvesant Town	166.056	45	1500	96200	1.0
10	East Village	186.245	10	3750	61700	16.5
11	Lower East Side	186.313	29	9999	44900	32.7
12	Upper East Side	189.268	10	7703	130300	3.3
13	Upper West Side	207.731	10	10000	119500	7.2
14	Murray Hill	221.277	0	2000	121200	1.6
15	Little Italy	222.066	41	5250	85500	16.8
16	Gramercy	223.211	45	3000	100400	4.6
17	Financial District	225.491	12	3000	124300	2.6
18	Chelsea	249.779	12	6800	98100	8.7
19	Greenwich Village	263.321	43	6000	120700	2.3
20	West Village	267.575	50	4000	124200	3.2
21	Midtown	282.784	30	5100	126800	9.6
22	SoHo	287.277	10	3000	121400	3.8
23	NoHo	297.855	75	1795	138900	2.4
24	Flatiron District	341.925	65	2000	129500	4.3
25	Battery Park City	367.557	55	7500	164700	1.4
26	Tribeca	490.638	60	8500	193900	3.6

4 - price, income, food stamps rate

```
/* STA402: Final Project
   Author: Karen Gaither
   Purpose: This project looks at the incomes in each neighborhood
            of Manhattan to determine if it has an effect on the
            prices of Airbnbs in those neighborhoods.
*/

%let wd = M:\STA402\Final Project;
%let bnb = AB_NYC_2019.csv;
%let income = manhattan_incomes.csv;
%let stamps = foodstamp_rates.csv;

* Output to an rtf file;
ods rtf file = "&wd\proj_output.rtf"
      style = journal bodytitle;

* Full Airbnb data;
data listings;
      infile "&wd\&bnb" firstobs=2 dsd;
      input id name :$55. host_id host_name :$35. borough :$20.
            neighborhood :$25. latitude longitude room_type :$16.
            price min_nights reviews last_review :$10. monthly_reviews
            host_listings availability;
run;
proc print data=listings (obs=10);
run;

* Drops Airbnb listings outside Manhattan and only keeps
   relevant variables;
data listings_manhat;
      set listings;
      if borough ne "Manhattan" then delete;
      keep name borough neighborhood room_type price;
run;
proc print data=listings_manhat (obs=10);
run;

* Median income in Manhattan neighborhoods data;
data income;
      infile "&wd\&income" firstobs=2 dsd;
      input neighborhood :$30. income;
run;
proc print data=income (obs=10);
run;

* Percentage of people on food stamps in each Manhattan neighborhood;
data foodstamps;
      infile "&wd\&stamps" firstobs=2 dsd;
      input neighborhood :$30. stamps_rate;
run;

* Merges the datasets using an inner join so it only keeps
   neighborhoods that are in both;
proc sql;
create table airbnb_half as
      select listings_manhat.name, listings_manhat.neighborhood,
             listings_manhat.price, listings_manhat.room_type, income.income
```

```
      from listings_manhat, income
      where listings_manhat.neighborhood = income.neighborhood;
quit;

* Uses another inner join to merge the food stamps data;
proc sql;
create table airbnb as
      select airbnb_half.name, airbnb_half.neighborhood, airbnb_half.price,
             airbnb_half.room_type, airbnb_half.income,
             foodstamps.stamps_rate
      from airbnb_half, foodstamps
      where airbnb_half.neighborhood = foodstamps.neighborhood;
quit;
proc print data=airbnb (obs=10);
run;

* Scatter plots of price vs. income of the different room types;
proc sgplot data=airbnb;
      scatter x=income y=price;
      where room_type = "Private room";
run;
proc sgplot data=airbnb;
      scatter x=income y=price;
      where room_type = "Shared room";
run;
proc sgplot data=airbnb;
      scatter x=income y=price;
      where room_type = "Entire home/apt";
run;

* Looks at the mean price in each neighborhood;
proc means data=airbnb;
      class neighborhood;
      var price;
      types neighborhood;
      output out=avg_prices mean=avg_price min=min_price max=max_price;
run;

* Merges the average prices data with the income data to see
   the values side by side;
data averages;
      merge avg_prices income foodstamps;
      by neighborhood;
      drop _type_ _freq_;
run;

proc sort data=averages;
      by avg_price;
run;
proc print data=averages;
run;


* Linear model with room type and income;
proc glm data=airbnb plots(maxpoints=20000)=diagnostics;
      class room_type;
      model price = room_type income / solution;
```

```
      output out=resid p=yhat r=residual;
run;
quit;

* "Residuals vs. Fitted Values Plot";
proc sgplot data = resid;
      scatter x=yhat y=residual;
run;

* Log transformation to deal with nonconstant variance;
data airbnb;
      set airbnb;
      log_price = log(price);
run;

* Fitted model with the log transformation of price;
proc glm data = airbnb plots(maxpoints=20000)=diagnostics;
      class room_type;
      model log_price = room_type income / solution;
      output out = logresid p=yhat r=residual;
run;
quit;

* Residuals plot with log transformation;
proc sgplot data = logresid;
      scatter x=yhat y=residual;
run;


* Scatterplot of food stamps rate vs. Airbnb price;
proc sgplot data=airbnb;
      scatter x=price y=stamps_rate;
run;

 * Tests if the food stamps rate is a significant predictor of price;
proc glm data=airbnb;
      model price = stamps_rate / solution;
run;
quit;



ods rtf close;
```