AirBnB - An Analysis of Consumer Surplus

Christina Ho, Nathan Gollogly

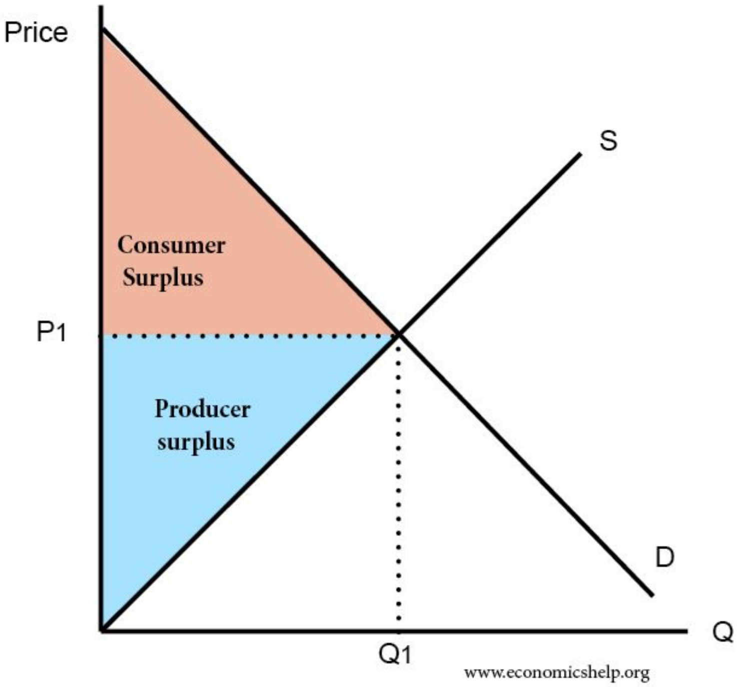
Business Problem

Our business problem is centered around AirBnB -- a platform business that provides and guides an opportunity to link two groups – the hosts and the guests. Airbnb gives hosts an easy way to monetize a space that would otherwise be going to waste.

Regulation of sharing platforms such as AirBnB has been prominent in the news over the past few years. After several years of litigation, AirBnb agreed to share data on user listings with NYC officials, making enforcement of the regulations possible[1].

In this project, we are acting as representatives of AirBnB when discussing regulation of the platform with government entities. We would like to show regulators (in an actual dollar amount) how much welfare the platform generates for the market we are operating in. In economic terms, this "welfare" is known as Consumer Surplus. Consumer Surplus is defined as "a measure of the additional benefit that consumers receive because they're paying less for something than what they were willing to pay."[2]

Consumer Surplus Diagram:



Our hypothesis that we plan to test is that AirBnB generates sufficient societal welfare (Consumer Surplus) such that any regulation which restricts Supply will have a negative impact.

[1] https://gizmodo.com/airbnb-will-help-nyc-track-down-illegal-rentals-and-ot-1835004746

[2] https://www.investopedia.com/terms/c/consumer_surplus.asp

Since this isn't a business context data science project and more of an economics project, there isn't much we say about how this problem is dealt with currently. With a data-driven approach, not only can this aid regulation but can also help the hotel industry revitalize their pricing algorithms especially during seasons of high demand. This could also potentially help hosts accurately help price properties using a wide range of data points. Paid third party pricing software is available, but hosts are generally required to put in average nightly price or base price.

As a subsequent phase of this project, we would seek to estimate the effects of regulation on supply and the calculate Deadweight Loss. Regulation would act like a "tax" on society and would (in theory) decrease Consumer Surplus.

Technical Setup

The data[3] for this project comes from a third party system that scrapes AirBnB listings, availability, and reviews and consolidates the data into 3 separate files. The data does not include information whether a listing was actually rented or not, however we are able to interpret this by using the presence of a review. These are the different data files we used:

listing_csv.gz: contains detailed listings data for New York showing 96 attributes for each of the listings.

calendar_csv.gz: contains detailed calendar data for listings in New York

reviews_csv.gz: contains review data for listings in New York

In order to calculate Consumer Surplus, we need 3 elements. The Price a consumer pays for a rental, the aggregated Quantity Demanded of AirBnB listings, and what consumers would have been Willing to Pay for a listing. The difference between what a consumer actually pays and what a consumer is willing to pay is that consumer's surplus.

Price and Quantity Demanded came directly from the data, however Willingness to Pay would have to be derived from the slope of the demand curve. Our first step was to attribute changes in Quantity Demanded to Price, not because an entire home in Manhattan is more desirable than a shared room in Queens. One cannot compare those listings to each other directly, so "clustering" the listings into like products would help us accurately compare similar listings to each other, in order to isolate the effect of Price.

In terms of initial exploratory analysis, we plotted the price trends to determine seasonality. Tuesday and Wednesday are on average the least expensive, both are around Monday prices. A high level of seasonality is evident, with a notable peak over the holiday period and on weekends.

The chosen four features of our data are: neighbourhood_group_cleansed (borough of listing), review score rating, room_type (hotel room, shared room, private room, entire home/apt), and host_identity_verified (true/false). We chose these features by reviewed histograms of many different features in the listings dataset settled on neighborhood group

[3] http://insideairbnb.com/get-the-data.html

(borough), room type, and host verification as our categorical data for modeling. This was because these categorical features were consistently present for each listing (few or no NULLS), and had a distribution that was not over-represented by a single value. We also included the Review Score as a numerical feature based on our own experience as consumers.

The clustering model we used was k-modes. K-modes is a frequency based approach to cluster categorical data (ie is the listing a shared or private room, what is the location, etc). These elements have no measurable “distance” between them, so overlapping frequency of similar elements allowed us to group like listings together.

Our data preparation involved “one hot encoding” our categorical data, and converting our numerical data (review score) into categorical data if it was above/below the mean. This process turned our listings data into dataframe of binary variables from which we could extract the values for the k-modes clustering. One of the challenges we faced was that we were unable to evaluate the performance of the clustering algorithm as there were no “known” values to test our data against.

From the k-modes clustering, the AirBnb listings were now segmented into “like” products from which we could plot Price and Quantity Demanded. Using regression analysis of Price and Quantity Demanded, we would be able to derive the slope of the demand curve needed to determine Consumer Surplus.

Price was the easiest element to derive from the data, as it was provided for each day a listing was available. To calculate for our time period, we simply took the average price per listing. Quantity Demanded had to be imputed based on the number of reviews a listing had. While not a direct indicator of whether a listing was rented or not, we felt it was a good proxy for demand. We also had to account for seasonality and availability since not all listings were available in equal amounts. This was accomplished by calculating a demand percent for each listing (the number of reviews a listing had ÷ days available) over the time period of our analysis (September 2018 - August 2019). This calculation was then annualized (multiplied by 365) to calculate the annual demand in days per listing. This ensured we were not over estimating demand for listings that were available every day of the year.

Each cluster was associated with the Price and Quantity Demanded data. Using Linear Regression, we the data was plotted for each cluster along with the slope, r-values, and intercepts. We were now ready to interpret our findings.

Results/Evaluation

Our k-modes clustering model performed as expected and segmented our overall AirBnB market into 6 distinct groups (we settled on k=6 after trying several versions of the model). We plotted histograms of the clusters and found clean sets of features that we were able to “interpret” into easy describe market segments. Our segments are as outlined below:

- **Segment 1: Value Private (All Boroughs)**
  These listings were from all boroughs with higher than average review scores, consisting primarily of the entire apartment/home, but where the host has not been verified.

- **Segment 2: Value Shared (All Boroughs)**
  This segment was similar to segment 1, but with listings only for rooms in an apartment shared with the host.

- **Segment 3: Budget (All Boroughs)**
  The listings in segment 3 were a mixed bag of below average ratings, primarily a single room in a shared apartment, with the host often not verified.

- **Segment 4: Premium (All Boroughs)**
  This was our premium segment, with the listings almost all for the entire home in Manhattan and Brooklyn, above average ratings, and host identity verified.

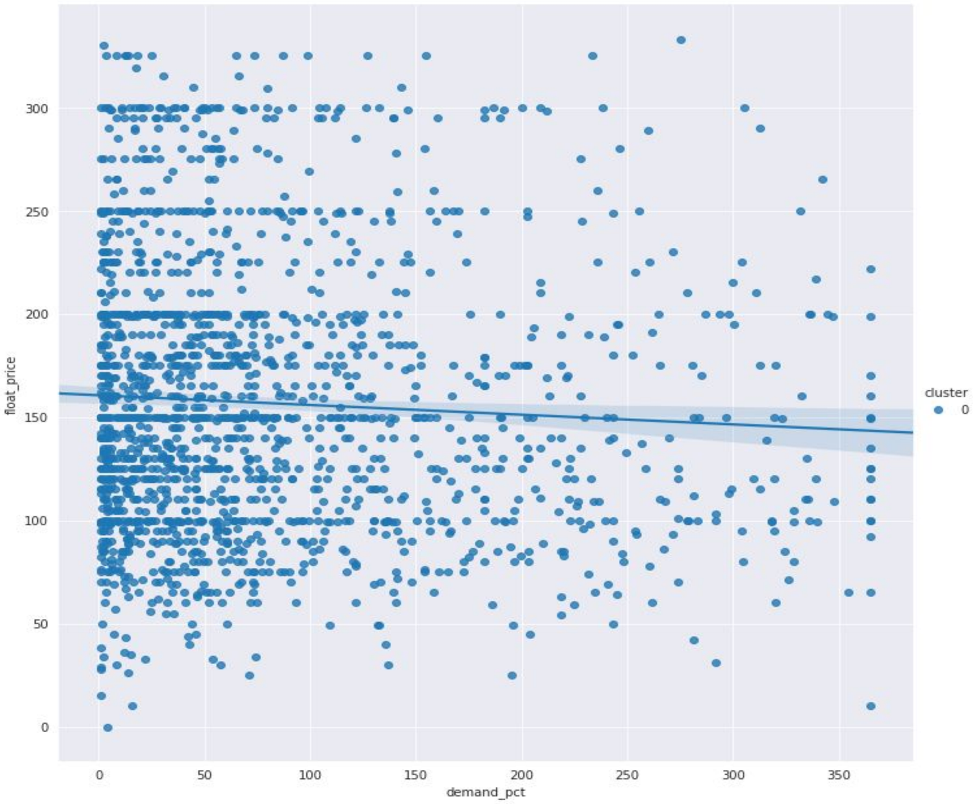
- **Segment 5: Premium Shared (excl Brooklyn)**
  Segment 5 was similar to 4, with high reviews and hosts verified, but these rooms are all in apartments shared with the tenants of the apartment.

- **Segment 6: Budget (Brooklyn)**
  Listings with below average ratings and listings for single rooms in shared apartments.

After we created our clusters, we were very excited to see how the dynamics of Price and Quantity applied to each segment. We expected our “Premium” markets to have steeper demand slopes (inelastic demand) and our “Budget” markets to be very responsive to Price changes.

We plotted Quantity Demanded (adjusted) and Price and ran a regression analysis on the results. However, the graphs we created and the statistical outputs did not match what we were expecting:

Scatter Plot: Segment 1 - Value Private (All Boroughs)
Slope: -0.0269 Intercept: 151.61 r_squared: 0.0011



Our "best" cluster (the cluster with the steepest demand slope and largest r-sq) was Segment 1 “Value Private”. This segment had a very gradual downward Slope of -0.03 and an R-sq value of 0.0011. This says that less than 1% of the variance in the dependent variable (Quantity Demanded) can be explained by a change in the independent variable (Price). Essentially, there is next to no correlation between Price and Quantity Demanded in this market.

Our clustering model and regression analysis did not match our hypothesis. We were able to calculate Consumer Surplus by calculating the average price for each segment and calculating the triangle between the average price, the intercept and the slope. For “Value Private” (our segment with the most promising results), we calculated a paltry Consumer Surplus of $79.20 for the year we were measuring. Further results included in appendix. The impact of regulation of this market would not be conclusive in terms of reduction to societal welfare (reduction in Consumer Surplus).

However, we felt there were takeaways for AirBnB. Having almost no Consumer Surplus may actually be a positive outcome from the company, as it means there is very little money being “left on the table”. Converting Consumer Surplus into revenue is a challenge many companies seek to achieve. Additionally, insight can be gained from having a clearer picture of the various market segments in the overall NYC AirBnB market.

Challenges and Changes

We were struggling to choose the most reasonable features to use in our k-modes model. Feature selection was especially important in this model because the listings dataframe contained 96 features including the unique id associated with the listing. A lot of the features includes, number of bedrooms, and bathrooms, amenities, review score, etc. Tests for feature importance were possible, but having mixed categorical and numerical features deemed to be problematic. First off, the k-modes model could only read in categorical data. Instead of means in the k-means model, the k-modes model matches dissimilarities for categorical objects.

If we were to do it again, we would use methods such as Chi Square tests and XGBoost for feature importance and statistical significance in using certain features. We could also run Principal Components Analysis for dimensionality reduction. We could do this by one-hot-encoding all the categorical features which also opens up possibly using k-means clustering. At the end, we ideally would want to compare the different clusters and dimensions we get to profiling different listings based on their features. In our current analysis, we are choosing features based on data cleanliness and our own domain knowledge of Airbnb as a customer.

Conclusion and Next Steps

Despite the results on Consumer Surplus being inconclusive, there are insights that can be gained from the analysis. The distinct market segments from the data clustering break down the overall New York City market into more discrete buckets. These market segments would help AirBnB target specific customers by creating unique and targeted marketing campaigns for each segment. Additionally, AirBnB can gain insights into what mix of criteria forms a higher overall demand by seeing which segments have the most Quantity Demanded and which have the least.

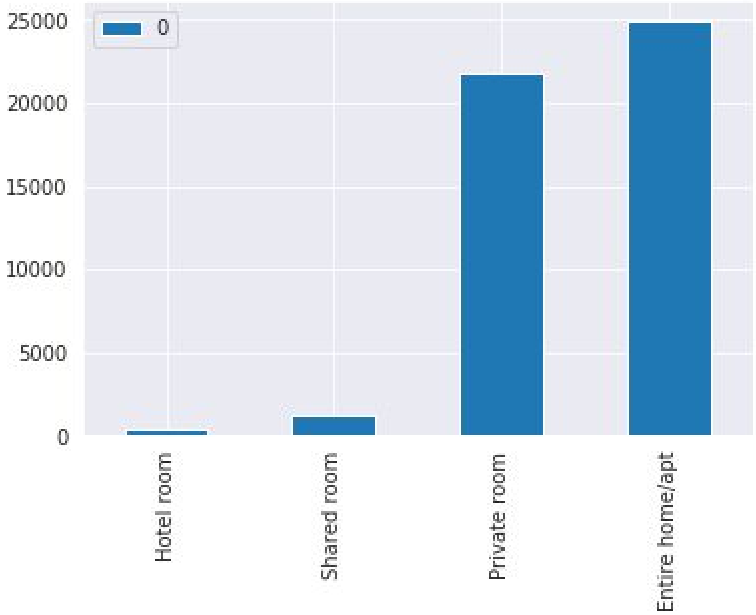
Had the results from the analysis shown a significant amount of Consumer Surplus in the market, AirBnB would have valuable insights when discussing regulation with government entities. There would have been a data driven approach for explaining the negative impacts of regulation. For next steps, the initial approach to estimate the impact of regulation on Consumer Surplus (deadweight loss), would not be applicable. However, AirBnB can still measure the effects of regulation by reviewing what segments are impacted. There may be unforeseen consequences that AirBnB can evaluate, such as whether the market shifts away from smaller neighborhoods in the boroughs outside of Manhattan.

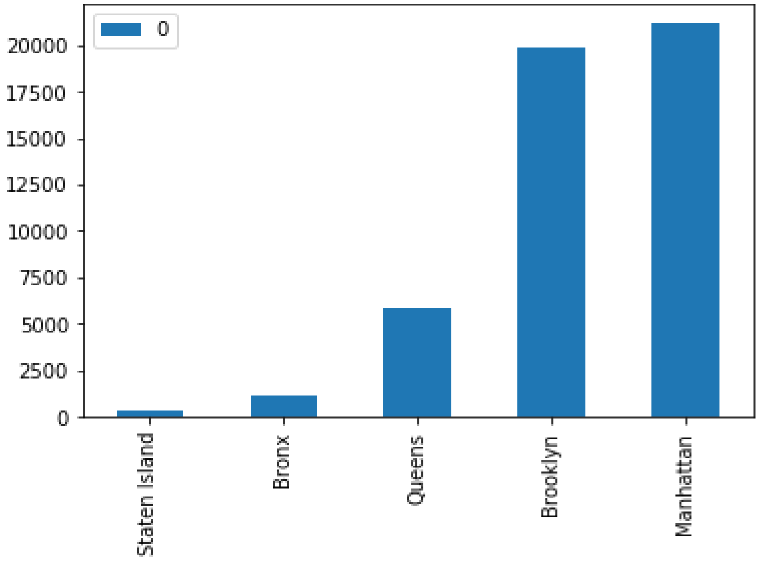
Lastly, AirBnB gains the insight that the market for rentals is driven by many factors outside of price. The sharing economy is a complex new model where consumers are driven by a myriad of factors when making decisions. Price will always be a component, but there are other factors that dampen the impact of Price when we choose where to stay when visiting a city. Exploring what these factors are will benefit AirBnB as a major player in the market for lodging and homeshares.

Appendix:

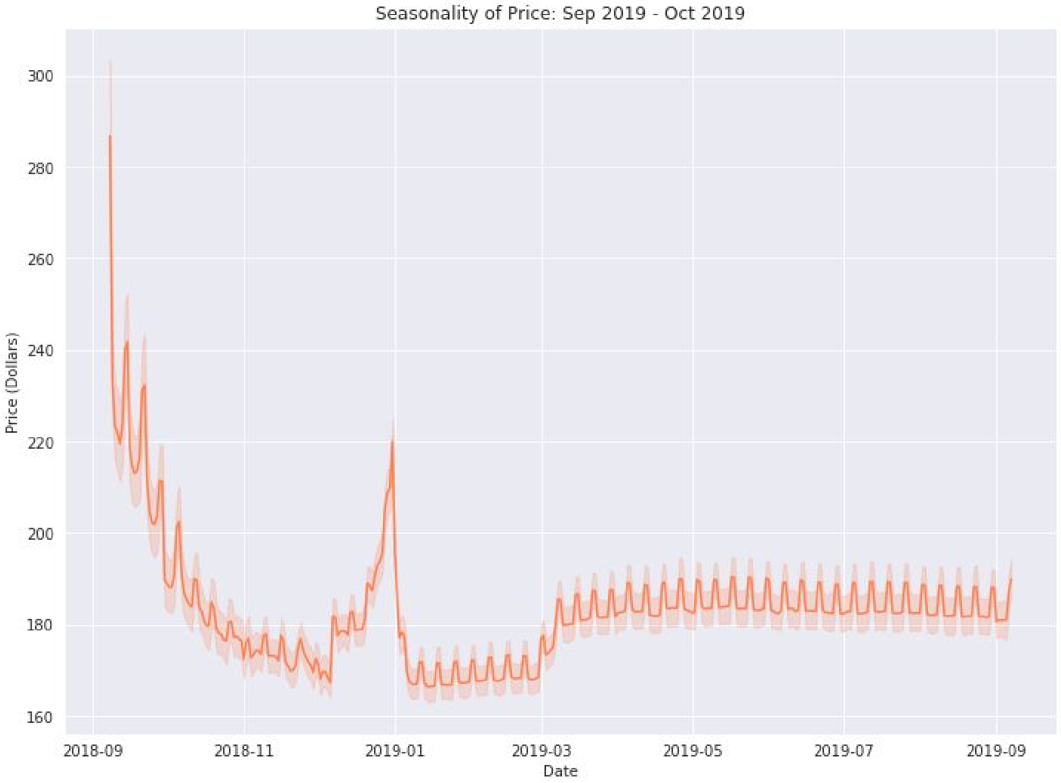
The workbook with our code can be accessed via this public link on Google colab

AirBnB Listings by Listing Type and Borough (Fig.1)

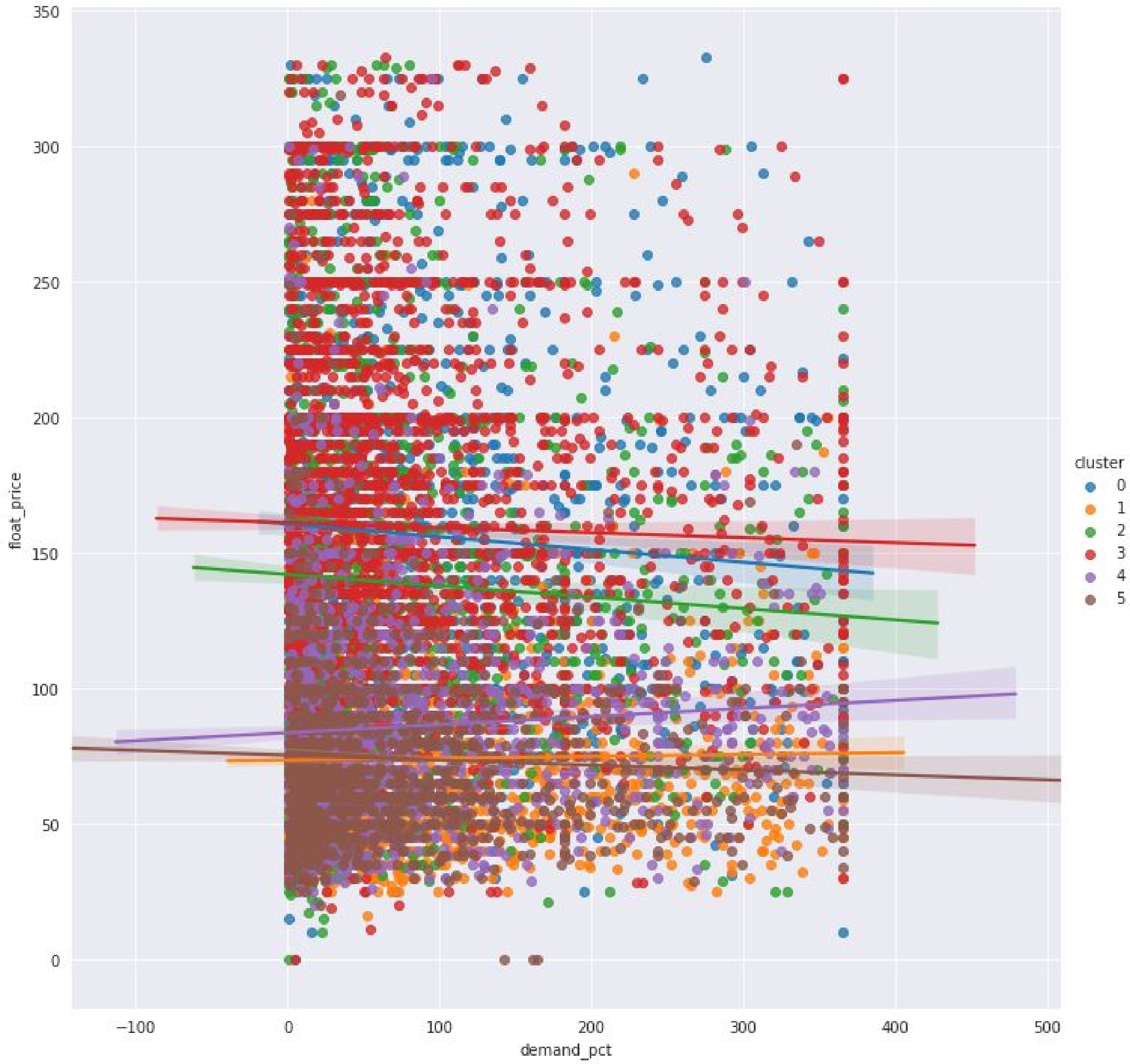




Seasonality (Fig.2)



Scatter Plot: All Segments



Statistical Outputs

Cluster 0 (Segment 1)

```
Slope:  -0.0269
Intercept:  151.6108
r_value:  -0.0339
r_squared:  0.0011
p_value:  0.1334
std_err:  0.0179
```

Cluster 1

```
Slope:  0.0068
Intercept:  73.3565
r_value:  0.01562
r_squared:  0.0002
p_value:  0.4389
std_err:  0.0088
```

Cluster 2

```
Slope:  -0.0134
Intercept:  135.2113
r_value:  -0.0148
r_squared:  0.0002
p_value:  0.4833
std_err:  0.0191
```

Cluster 3

```
Slope:  0.0171
Intercept:  150.7101
r_value:  0.0205
r_squared:  0.0004
p_value:  0.2361
std_err:  0.0144
```

Cluster 4

```
Slope:  0.0299
Intercept:  83.4432
r_value:  0.0573
r_squared:  0.0032
p_value:  0.0183
std_err:  0.0126
```

Cluster 4

```
Slope:  -0.0179
Intercept:  75.1434
r_value:  -0.0432
r_squared:  0.0018
p_value:  0.0967
std_err:  0.0108
```