Visualization of massive data

Does the price of an Airbnb at New York influence the length of the stay?

Dataset: AB_NYC_2019.csv https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Description: "AB_NYC_2019" - Summary information and metrics for listings in New York City. It is good for exploration, visualizations and predictions.

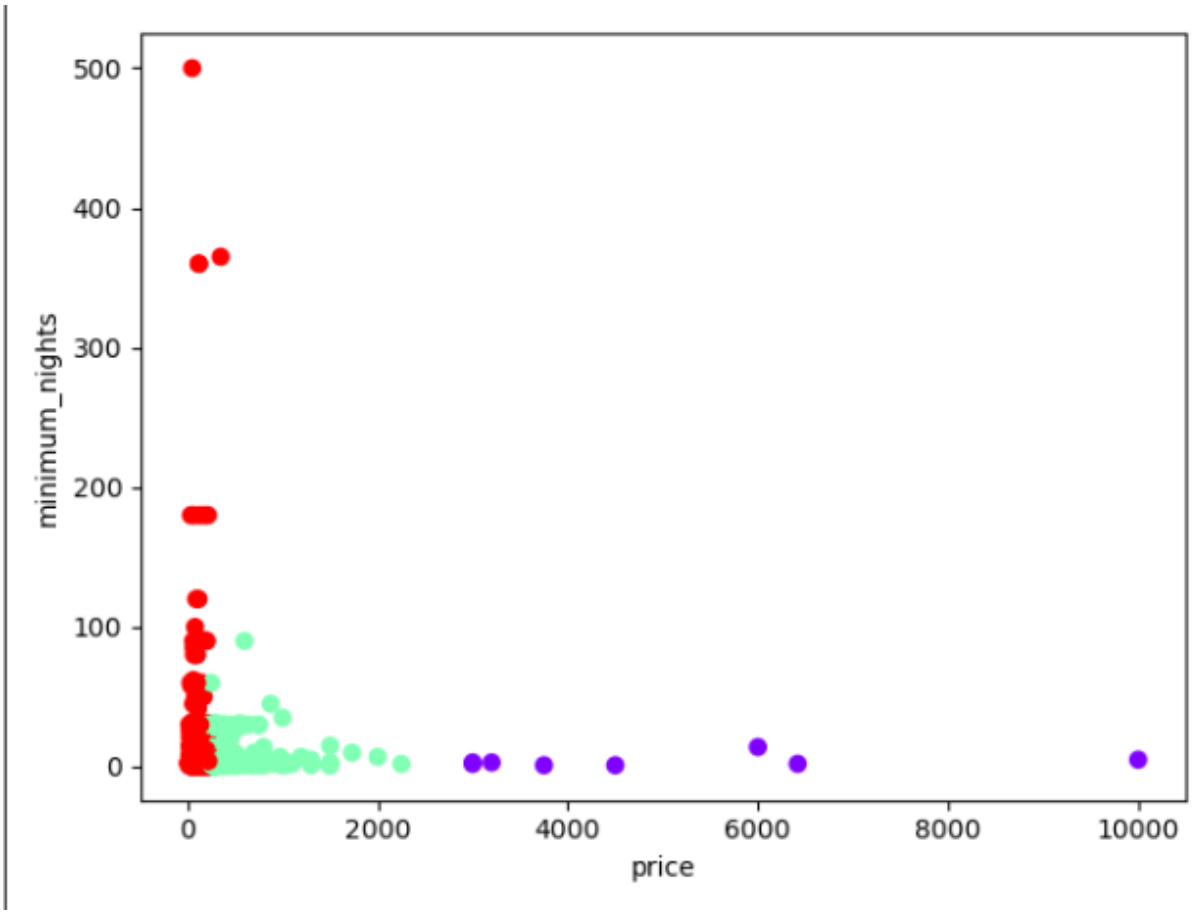
Columns (16 total):

- **Id**: listing ID
- **Name**: name of the listing
- **host_id**: host ID
- **host_name**: name of the host
- **neighbourhood_group**: location
- **neighbourhood**: area
- **latitude**: latitude coordinates
- **longitude**: longitude coordinates
- **room_type**: listing space type
- **price**: price in dollars
- **minimum_nights**: amount of nights minimum
- **number_of_reviews**: number of reviews
- **last_review**: latest review
- **reviews_per_month**: number of reviews per month
- **calculated_host_listings_count**: amount of listing per host
- **availability_365**: number of days when listing is available for booking

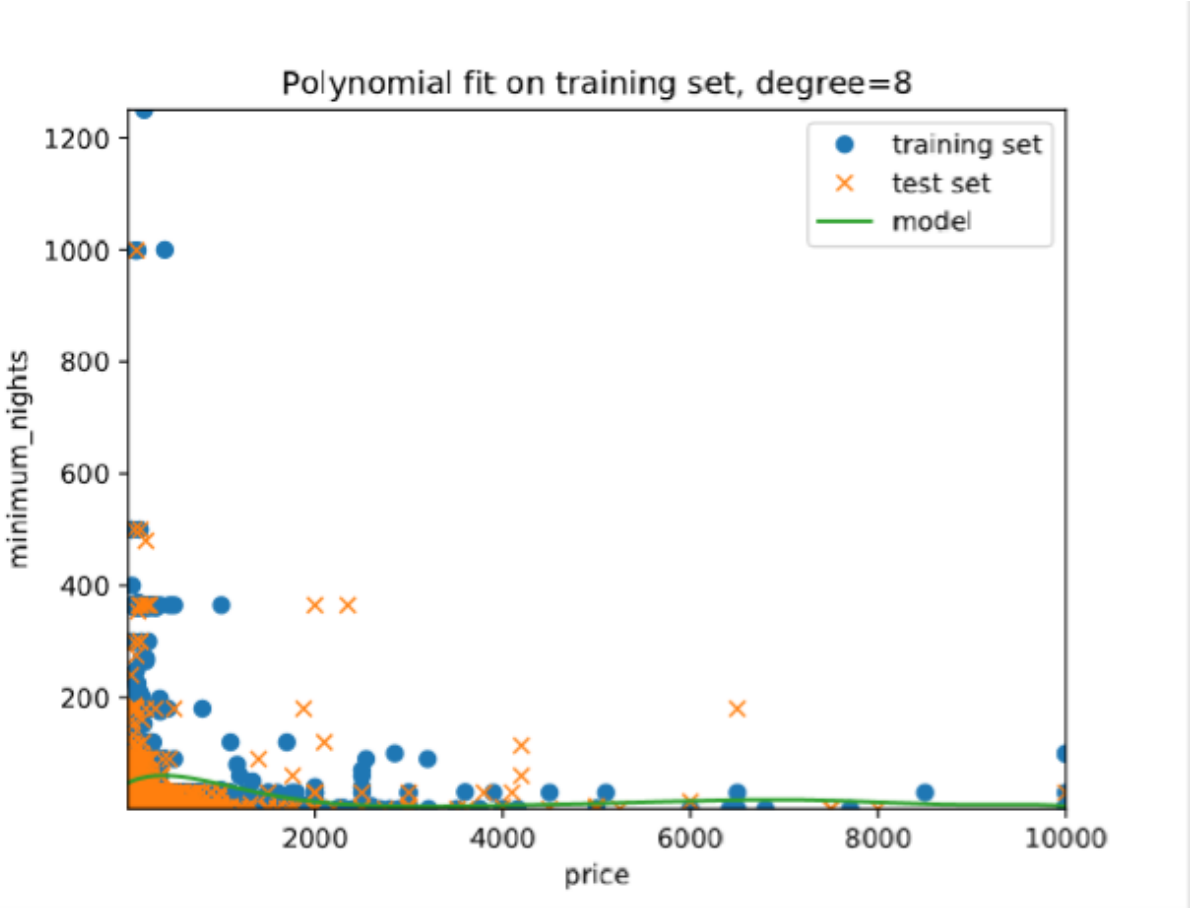
Correlations between variables:

Following our question, the hypothesis we can formulate is that when the price column has higher values, the minimum_nights column would have lower values. We could go further and make a hypothesis correlating the availability_365 column with the two previous columns seeing as more expensive Airbnb's could tend to be holiday houses and so would only be available for booking during school holidays when the owners aren't living in the home. This would in turn also diminish the number of minimum nights possible for said Airbnb.

To be able to plot a relatively correct graph, we first had to process the data. The first step for this was to eliminate null values and clear the data set. Then when visualizing, the first choice of plot we went for was a scatter plot because it allows us to clearly see the correlation between the price and the minimum number of nights taken. Once the data was plotted, we could clearly see a tendency for more expensive Airbnb's to be reserved for shorter times contrary to cheaper Airbnb's.



After plotting the results, we can see that the higher the price goes the less likely there is to be a reservation and a long stay. The red dots which represent the low prices can go up to over a year of reservation however as we go further up in the price with the green dots and the purple dots, we can see a general trend for stays to be shorter. This can be attributed to the fact that more expensive Airbnb's tend to be used for holidays instead of student accommodation for example for cheaper Airbnb's.



For the second part of this project we chose to create a predictive model of a column as a function of another column. As we can see in the picture above, the model was trained with a degree of 8 and fits quit well the data set thereafter. We can see a slight bump in the curve between the 6000$ price and 8000$. We theorised this was caused by the fact that the degree was either slightly too high or the training set's random values gave us pricier Airbnb's in that region of the graph.