Project I Data Story
Anna Apa

After the first step of collecting and cleaning the datasets I needed, I started looking into any possible trends between the features of the Airbnb listings.

Grouping by Neighborhood and Studying Trends Using Barplots
Cancellations, Listings, Reviews per Neighborhood

To see the distribution of the Airbnb listings across the London neighborhoods, I created a barplot that plots the total listings per neighborhood. We can see that the distribution of the listings between the neighborhoods are widely varied. We see that Westminster, Tower Hamlets, Hackney, Camden and Kensington and Chelsea have the most Airbnb listings (as confirmed by the 'top 5' table created).

I then do this with the number of reviews and number of cancellations as well. I see that Westminster and Tower Hamlets are still the top two neighborhoods who have the most in both aspects. Could this mean that listings in their neighborhoods have the highest likelihood of cancelling or is it just because they also have the most number of listings? I try to answer this question by creating a barplot that shows the ratio of the number of cancellations to the number of listings per neighborhood. This barplot shows us that 1) the distribution is no longer as spread out and 2) Westminster does not have the highest ratio.

Days Booked vs Cancellations Scatterplot

Is there any relationship between the popularity of a listing (more booked days) and its history of cancelling on guests? By creating a scatterplot between the two variables, we see that there is no trend which tells us there isn't a strong correlation between the two.

The Relationship of Price with Other Variables

First, I study if the number of cancellations vary among listings of different prices. In the scatterplot produced, I see that there seems to be a negative exponential correlation between the two.

It is also important to note that the price of an Airbnb listing varies by room type. In this dataframe, we see that there are three room types in the Airbnbs in London: entire home/apt, private room and shared room. I create a pivot table and a heatmap to see which neighborhood has the most and least expensive listings per room type. This

heatmap shows us that an entire home/apt room type tend to be the most expensive(as expected) and the City of London has the most expensive listings.

I play around more with heatmaps to visualize the pivot tables I made to better understand the trends between these features better.

For the second heatmap, I check to see which neighbourhood has a history of cancelling the most, by room type. We see that Kensington and Chelsea has the highest history of cancelling on entire home/apt room type listings while Tower Hamlets has the highest number of cancellations on private room type listings.

Taking Into Account Crime Rates

I create a heatmap that shows the crime rate trends of each neighborhood over the years. We can see that crime rate has lowered down a lot since 1999. We also see that although its crime rates have gone down, Westminster still has the highest crime rate among the London neighborhoods.

Since we're mainly concerned about the crime rate history of the neighborhoods starting on the year the Airbnb data we have was collected, we only look at the crime rates of each neighborhood from 2011-2017. We see that Westminister's crime rate is not as drastically different from the other neighborhoods as before.

Is there a relationship between the neighborhood's crime rate and its listings number of cancellations? By creating a Crime Rate vs num_cancellations scatter plot, we see that there is yet another random relationship between the two variables which means there is probably no relationship.

Does the crime rate affect the pricing? We create a swarmplot to study this relationship but again, the trend is random.

Now that we have a better idea of the relationships of the variables with one another, we have now potential hypotheses we can explore in the EDA report for this project. In this project, we can see the correlation and possible causation of cancellations with variables such as the room type, neighborhood, crime rates, price and etc.