When I first reviewed the Kings County data, one of my first considerations was whether or not the final regression model should use only intrinsic measures to predict the values of homes. By this, should the model only consider the number of bathrooms that a home has, or should factors such as its zip code be included? With this being my first data science project, the idealist in me felt that intrinsic predictors should only be part of the final model. However, as I became more involved in the project, I began to see the value of considering measures that take in local market dynamics such as the square footage of homes in that neighborhood when developing the final model. From a data science perspective, predictor variable selection is a very important process. However, it would be unwise not to also consider how a data scientist’s own lived experiences affect the importance they place on variables they encounter when working on a particular project.

The desire to only include characteristics of the home such as the quantity of bedrooms or square footage of living space is informed by some of my own lived experiences. As a native New Yorker, born and raised in the borough of Brooklyn, I am happy to see the surge in home prices in many of the neighborhoods I frequented as a child. From brownstones in Fort Greene to condos in Downtown Brooklyn, the real estate market in Brooklyn is very competitive today. Yet, this wasn’t the case when I was a high school student in the early 2000s. While brownstones were always priced nicely in Brooklyn, it was uncommon for them to get prices north of a million dollars as they now do today. Further, the dramatic increase in neighborhood institutions that buoy home values such as coffee shops is much more pronounced today than it was during my adolescence. Thus, if the features of the homes and their neighborhoods never changed, what is the cause of the market perceptions that have caused housing values to explode and Brooklyn become a destination to live in New York City? This personal area of interest around why certain communities do not enjoy appreciation in home values as comparable neighborhoods is something that I have been passionate about recently in the Atlanta-area.

This personal interest informed my initial work on the project by making me somewhat biased against certain predictor variables. On my first attempts to preprocess the data, I found myself spending more time making sure that the distribution of the bedroom column was as close to normal than the longitude variable. This wasn’t because I have some weird obsession with designing bedrooms as a hobby. Instead, my personal feelings on housing markets subtly influenced this behavior. Likewise, it wouldn’t have taken much to convince me that the zip code column could be withheld from inclusion in the final model. If the task for the project was to predict a home’s value, I did not see how including the home’s zip code was as valuable as knowing the number of floors that the home had for example. Again, my personal experiences and belief system, not empirical evidence, informed this way of thinking.

For the purpose of this project, personal biases shared by the data scientist do not have a major impact on society. However, reflecting on how I responded to the data when forming my initial impressions of it highlighted the importance of being a transparent practitioner. There will be future instances when we will examine other areas and phenomena that one might have personal feelings about that can affect analysis. I don’t believe that data scientists have nefarious motives when they allow their own biases influence their judgments at all. As humans, we generalize and stereotype everyday to help provide some framework from which we can make sense of the multiple data points we come across daily. Our fear of touching the hot tea kettle on the stove is informed from a past experience of touching something that is hot and feeling the pain of burning one’s hand. Likewise, when presented with a topic of analysis that we may have formed past conclusions about, we may subtly take actions to ensure that our past assumptions about a matter are confirmed in the presence of new data that might challenge our generalizations.

As it applied to the Kings County data, my first models did not take into consideration factors I felt were extrinsic at first. I overlooked the importance of considering how many times a home was viewed during the sales process and totally excluded zip code from the data set. As the data scientist, I was willing to justify these decisions with a nicely written statement that spoke to the value of knowing a home’s “true value”. However, when my coefficient of determination results routinely came back in the mid-40s, I knew that I was doing something incorrectly and had to challenge my initial stance because the summary results were not ambiguous to say the least. As I began to include more factors such as the condition of the home and whether it was near the waterfront, I saw improvement in my model’s output.

For the purpose of this project, such a realization around confirmation bias might not have much impact on the larger data science community. However, as a new data scientist, this project has made me personally resolved to championing diversity in thought moving forward. Failure to seek a multitude of perspectives can result in products being designed that do not meet the public’s need. One example of this in the real-world are self-driving cars that have difficulty identifying individuals on the streets that are people of color. In the design stage of this feature of the car, even if the product teams were not homogeneous in composition, these teams clearly exhibited some form of groupthink when building the program that identified pedestrians. If someone on the team would have realized that they were only selecting characteristics that identify people on the streets that reflect their own lived experiences, they would have been more inquisitive to learn how to identify people that were female or were of different races. Further, being self-aware about one’s own biases can ensure that the tests put in place to verify an item’s validity do not simply confirm one’s biases.

Diversity has become a buzzword in recent years and often is used to bring attention to the historical lack of representation in fields such as technology. While it is important to strive for more women and people of color in fields such as data science, the argument can be made that it is just, if not more valuable, for data scientists to challenge their own personal biases when engaged in analysis. If more data scientists possess the skill to self-report when personal judgments can skew their results, they will design tests and other measures that help challenge preconceived notions. Unlike the summary chart from the Kings County project that called attention to my bias around excluding factors I thought were market dominant, many projects in the future might be difficult to identify and can go by unchecked. As I move forward on this journey to becoming a data scientist, I will be sure to track how I am able to compartmentalize personal biases I might have on subjects of study so that it doesn’t influence decisions made when conducting various analyses.