**Summary of Visualization**

My visual shows that members of more affluent households ($75,000 per year and up) clearly are more likely to partake in at least 10 minutes of recreational physical activity, moderate or vigorous, in a typical week. Furthermore, it shows a more nuanced picture of how physical activity becomes more likely with increasing income. For instance, it shows that vigorous physical activity very suddenly becomes more likely to be reported when moving from the $20,000–$34,999 income bracket to the $35,000–$54,999 bracket. There is much room to speculate on why this trend exists (e.g., the lack of safe and open park space in lower-income neighborhoods).

**Data Sources**

The data are all from the CDC’s 2015-16 NHANES survey. The household income data are part of the survey’s [demographic dataset](https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2015). The physical activity data are part of the [questionnaire dataset](https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2015) (under PAQ\_I).

**Notes on Methodology**

I chose household income over other measures of income because I believe that it most accurately reflects the living situation one is in. For example, a stay-at-home parent in a multi-family home might have little to no individual income but live with their high-earning spouse and with a different family that earns much more than their own. Binomial 95% confidence intervals are shown on my graph to acknowledge the uncertainty inherent in my data source, which is a survey.

The dataset also includes personal estimates of how many days and how many minutes interviewees recreationally exercised in a typical week for each activity type. I went with the simpler yes-or-no question on whether one does more than 10 minutes of a physical activity type in a typical week because the other two metrics are somewhat prone to measurement error because of flaws in interviewees' memories. Furthermore, the standard deviations for both metrics are extremely wide, causing confidence intervals to extend well below zero. I demonstrate this in the third and final cell of my source code. The code outputs the last 14 columns of the data table off which the graph would be based, where all confidence interval upper and lower bounds are.

There are also several bins from the dataset that I did not include in this graph:

* "Refused"
* "Don't Know"
* "Missing"
* "$20,000 and over" - according to the NHANES documentation, an observation only goes into this bin if the interviewee is unwilling or unable to give an exact number for their household income, and they cannot give any more detail than whether their income was less than $20,000 or $20,000+. This bin did not fit with the way I categorized income, so I excluded it.
* "Under $20,000" - excluded for symmetry with excluding "$20,000 and over," even though it fit with the way I constructed my income bins.

In my code, I include a second part with an alternate data set where this category is assimilated into the "<$20,000" bin (compare keydatameans with keydata\_altmeans). The resulting summary statistics are little different from the data used for the graph. If anything, I am being conservative by excluding this bin because including it tightens the 95% confidence interval for the <$20,000 bin and slightly lowers the percentage of those in the $20,000 who report moderate physical activity (31.53% 🡪 30.88%). The percentage reporting vigorous physical activity in this category remains virtually the same (14.61% 🡪 14.64%).

Income bins are not all the same size (e.g., the $75,000 to $99,999 category covers $24,999, while the $20,000 to $34,999 category covers only $14,999). However, this is a product of the specific bins that the data itself had for household income. If I had used the original bins, there would be 12 in total (excluding bins that I omitted). I combined bins in a way that I thought minimized the disparities in bin size while ensuring that bins had a roughly similar number of observations in them. I grouped the original income bins as follows:  
"$0 to $4,999," "$5,000 to $9,999," "$10,000 to $14,999," and "$15,000 to $19,999" 🡪 "<$20,000"  
"$20,000 to $24,999" and "$25,000 to $34,999" 🡪 "$20,000–$34,999"  
"$35,000 to $44,999" and "$45,000 to $54,999" 🡪 "$35,000–$54,999"  
"$55,000 to $64,999" and "65,000 to $74,999" 🡪 "$55,000–$74,999"  
"$75,000 to $99,999" 🡪 "$75,000–$99,999"  
"$100,000 and Over" 🡪 "Over $99,999"  
  
The $20,000 mark is also particularly meaningful because it was the approximate location of the U.S. federal poverty line for a family of 3 in 2016. The $55,000 mark is also significant because it is close to the U.S. median household income ($57,167 in 2016, per the [U.S. Census Bureau](https://www.census.gov/library/publications/2017/acs/acsbr16-02.html). I chose to keep the "$20,000–$34,999" separate from the "$35,000–$54,999" category and the "$55,000–$74,999" separate from the "$75,000–$99,999" category because doing so preserves certain useful information. For example, we can see in the visual that the upper bound of the 95% CI for the "$20,000–$34,999" for vigorous activity is lower than the lower bound of the 95% CI for the next income bracket up, "$35,000–$54,999."