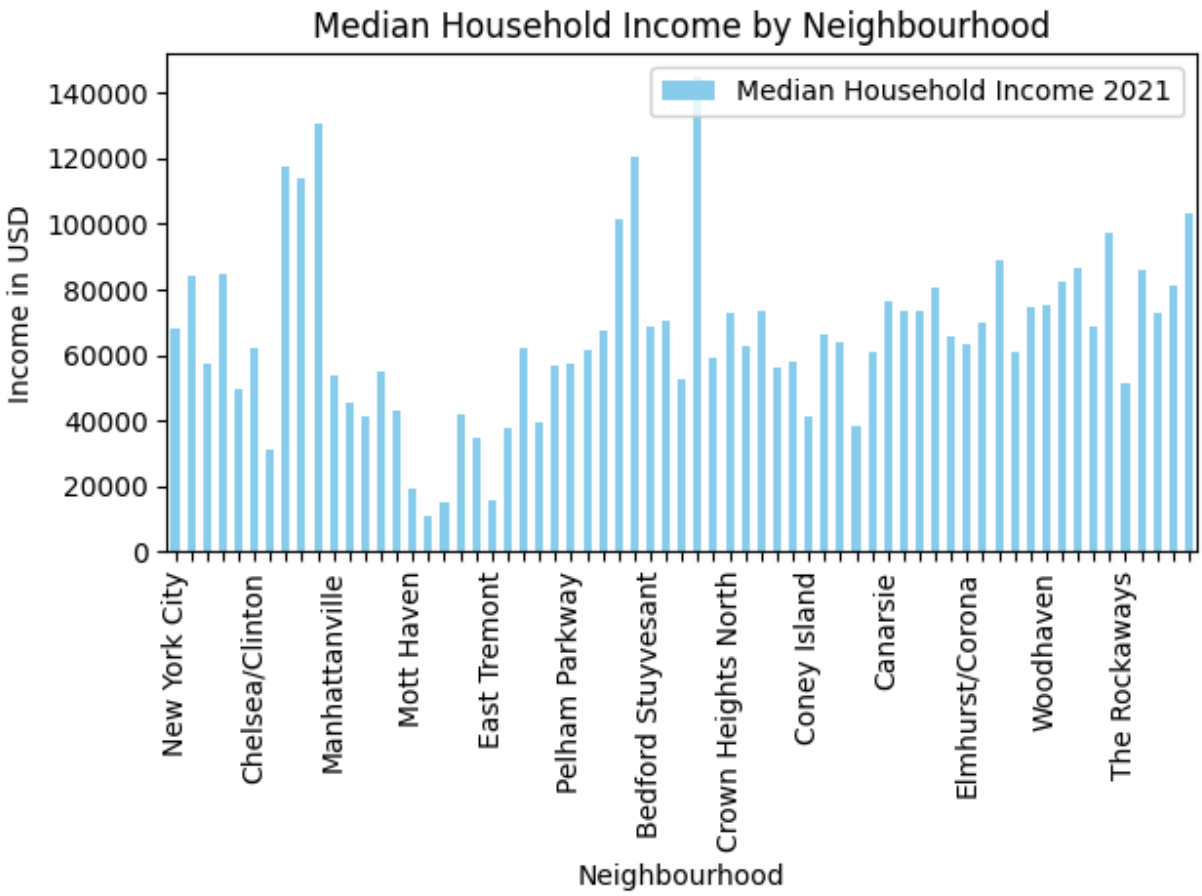
Visualizations

1. Median Household Income by Neighbourhood



- Why did you pick this representation?

  It is easiest to visually see the differences in wealth between neighborhoods by looking at a bar chart of the income distribution. From observing this graph we can easily see how neighborhoods like Manhattanville have much higher income than for example Mott Haven.

- What alternative ways might you communicate the result?

  We could use radial column charts to also effectively communicate disparities in wealth between different neighborhoods.

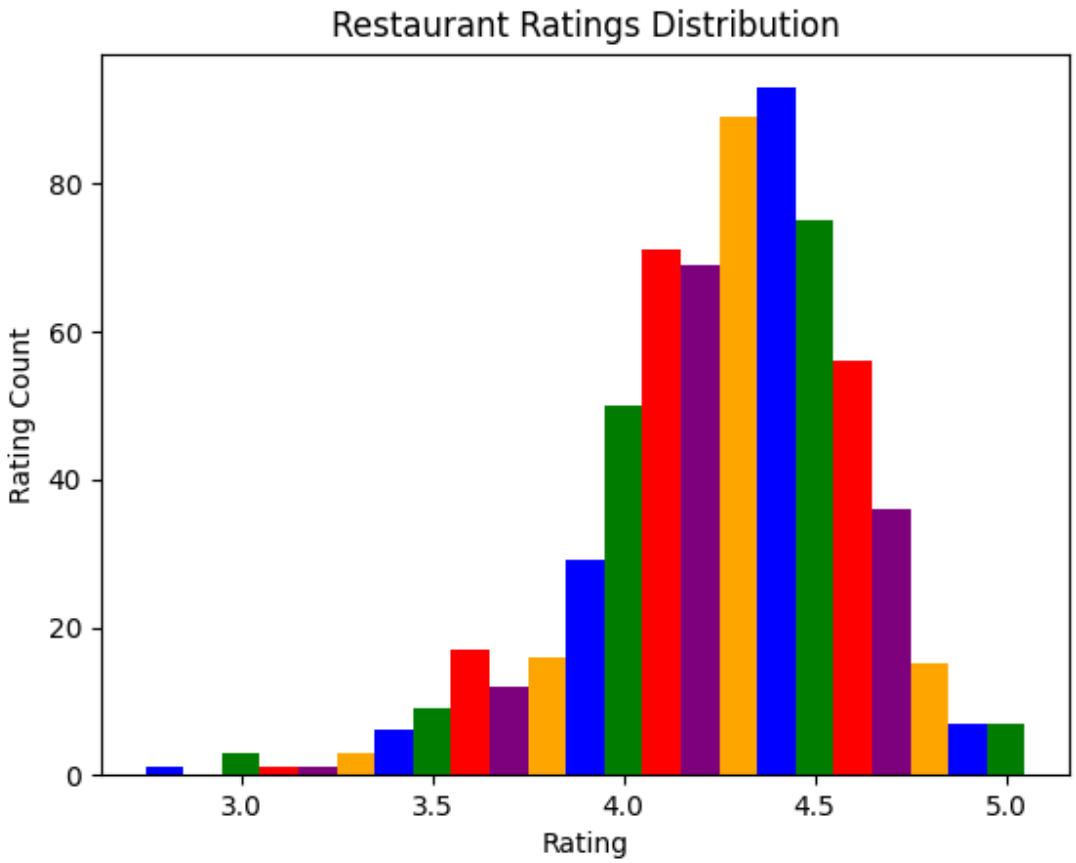
- Were there any challenges visualizing the results, if so, what were they?

  It was challenging to show all the neighborhood names on the x axis since they would overlap with each other, so we decided it was best to simply show every 5th name or so to make the graph cleaner. This means we lose the specifics about which exact neighborhood has what precise income, but the tradeoff for better readability is worth it.

- Will your visualization require text to provide context or is it standalone (either is fine, but it's recognized which type your visualization is)?

    No text is required, this visualization is readily understandable based on axes titles and the graph title itself.

2. **Rating Counts Distribution**



- Why did you pick this representation?

    This representation helps show the distribution of ratings amongst all the restaurants in our dataset. As you can see there is a much higher proportion of ratings between 4.0 and 5.0, with frequency of ratings outside this range falling off drastically. In fact we don't even see any ratings below a 2.5 and barely any ratings below 3.5.

    This helps demonstrate the almost psychological bias people tend to have in their ratings, such that a great concentration of ratings occurs around 4. This graph looks like a normal distribution skewed to the right with a mean in the 4.0-4.5 range.

- What alternative ways might you communicate the result?

    We could've used a pie chart with the different slices being either rounded whole number ratings or .5 increment ratings to show the proportion of total ratings they

account for. However we felt that in discretizing the rating range that much, we would lose some information regarding the distribution and thus we chose this histogram as it naturally felt like the best way to visualize this information on Rating frequency instead.

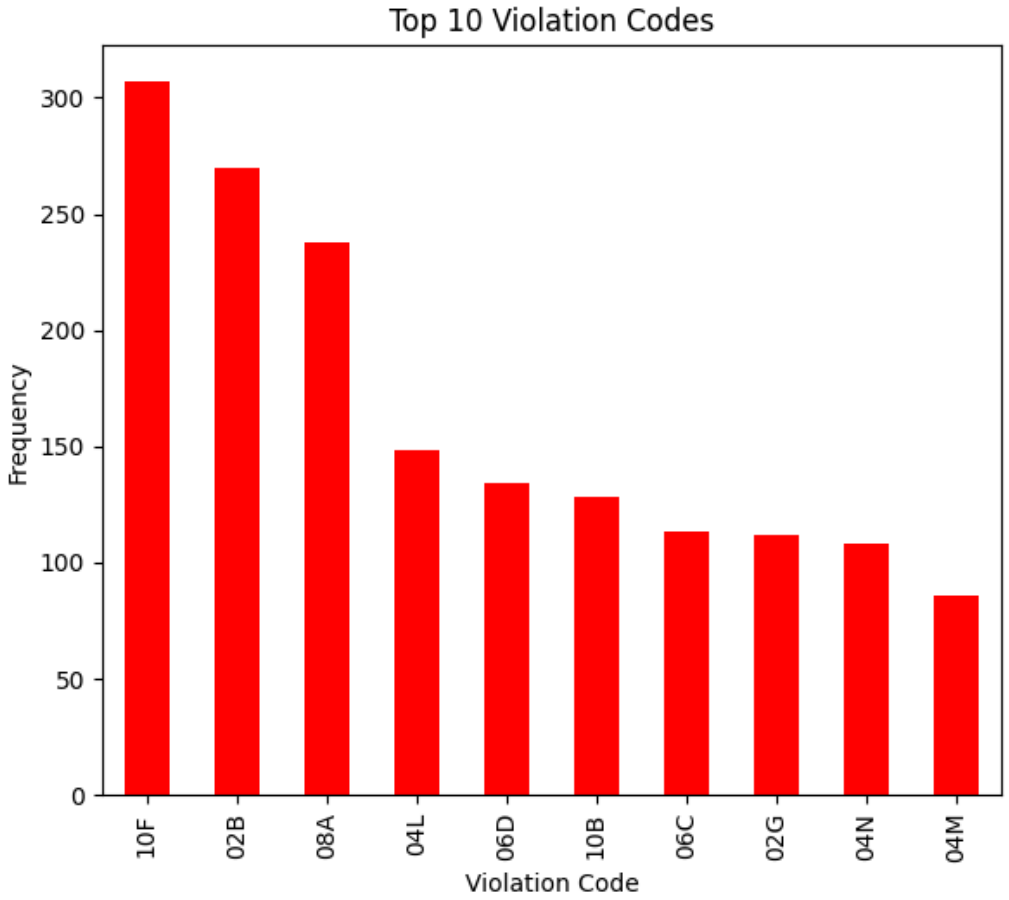
- Were there any challenges visualizing the results, if so, what were they?

  We had some challenges in deciding upon a discretization granularity. We eventually settled on 0.1 discretizations as this is the finest grain-level offered by our dataset, so it would not be losing any information in this representation. We considered using 0.2 or some lower granularity levels but we felt visually we were losing too much information about the distribution in doing so.

- Will your visualization require text to provide context or is it standalone (either is fine, but it's recognized which type your visualization is)?

  It only requires the context that these are Google ratings for restaurants in the NYC area. However after that it is pretty self-explanatory in its showing of the distribution of frequencies of each of the various possible rating scores out of 5.

3. Top 10 Violation Code Frequencies



- Why did you pick this representation?

This representation helps isolate the top 10 most frequent health code violations and gives a sense of scale as to how much of each occurs. We can tell that the codes 10F, 02B, and 08A have the most frequency. 10F corresponds to non-food contact surfaces such as floors, ceilings, equipment being improperly constructed or maintained. 02B corresponds to potentially hazardous foods (PHFs), which are foods on which microorganisms can grow and thus contaminate, being kept below 140F. 08A corresponds to the facility not being vermin proof, which is likely an indication of rats, insects, other creatures that have the potential to transfer diseases and make it a non-hygienic kitchen. This helps give a sense of what are the major more common issues facing restaurants in the NYC area.

- What alternative ways might you communicate the result?

  Here again we could've used a pie chart to convey proportions instead of absolute value of occurrences. We thought the absolute scale of exactly how many of these infractions were happening was important and that would be lost if we opted for a pie chart. Visually we can still get a rough perception of relative proportions through this graph.

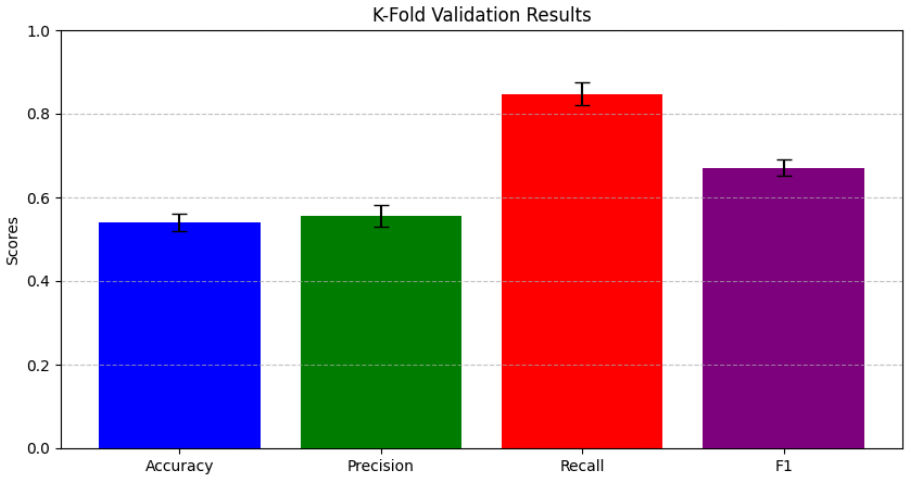
- Were there any challenges visualizing the results, if so, what were they?

We had trouble deciding how many of the code violations to show, but ultimately settled on 10 as they are an adequate representation of the bulk of code violations amongst NYC restaurants. Also having more than 10 made the graph visually overloading and provided too much for viewers to pay attention to. Top 10 struck a balance between being visually informing enough while not overwhelming.

- Will your visualization require text to provide context or is it standalone (either is fine, but it's recognized which type your visualization is)?

  Yes this would require text contextualization stating that the x axis were the health violation codes given to different restaurants in the NYC area. We would further require a map/legend between the error code and what their meaning is, such as I did above for the first 3 code violations, so that the layman can understand what these codes and frequencies represent in their restaurant frequenting and consumer context.

4. Logistic Regression Metrics



- Why did you pick this representation?

  We picked this representation to show and contrast our different metrics for the Logistic Regression model. We had 4 metrics with different error bars. You can see from the accuracy that the model performed with 50% accuracy and since this was a binary classification task that likely indicated that there wasn't as strong of a relationship between our feature variables and our target. High Recall suggests that most of the positive instances are classified correctly even if negative instances are falsely classified as positive. F1 indicates a measure of the model's overall performance incorporating Precision and Recall.

- What alternative ways might you communicate the result?

  There weren't any alternate ways to present this as this graph is just a means to compare and contrast the different metrics of our logistic regression model. Bar chart made the most sense here.

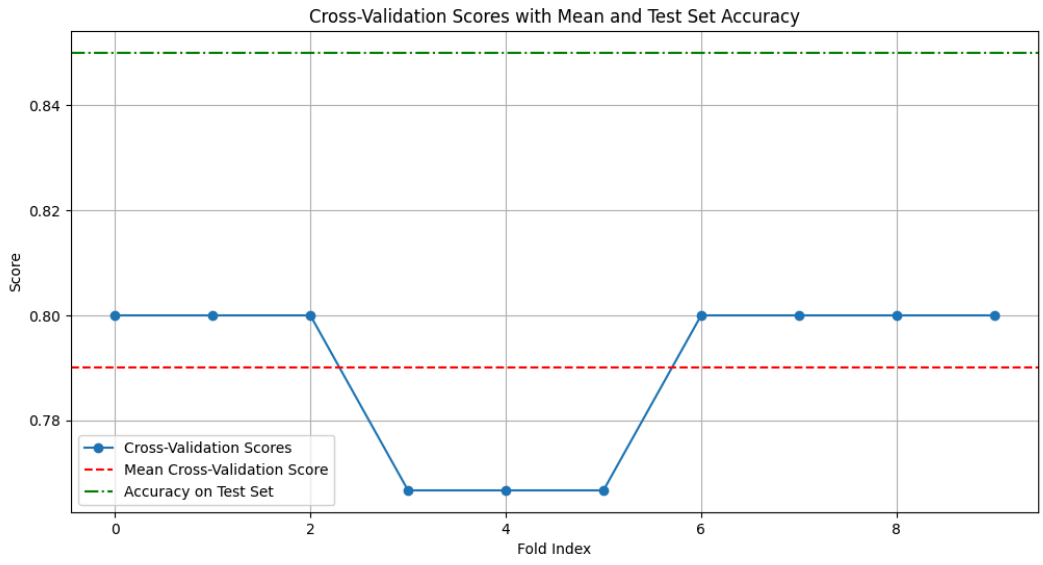
- Were there any challenges visualizing the results, if so, what were they?

  There were no challenges visualizing these results.

- Will your visualization require text to provide context or is it standalone (either is fine, but it's recognized which type your visualization is)?

Yes, we need to provide context about what each metric means so that lay people can understand what each value for each metric represents.

5. K-NearestNeighbour Metrics



- Why did you pick this representation?

This representation shows the different metrics we collected for our Knearest neighbour model. It shows the mean cross validation score and the variation of the cross-validation scores across each k-fold index. It also plots the average test accuracy as that is another metric we collected.

- What alternative ways might you communicate the result?

There were no alternative ways to compactly view all three of these results on one graph. This graph does it in a way that is not visually overwhelming and actually quite visually appealing.

- Were there any challenges visualizing the results, if so, what were they?

  There were no challenges visualizing the results.

- Will your visualization require text to provide context or is it standalone (either is fine, but it's recognized which type your visualization is)?

No context is required, it is a standalone visualization. The only context that may need to be provided is greater detail on how cross-validation scores are calculated in case lay people want greater understanding.