The following table lists the approximate populations and land areas of the five boroughs of New York City (data obtained via https://en.wikipedia.org/wiki/Boroughs_of_New_York_City#Background):

<table>
<thead>
<tr>
<th>Borough #</th>
<th>Area (square miles)</th>
<th>Population (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Bronx</td>
<td>42.1</td>
<td>1.4</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>70.8</td>
<td>2.6</td>
</tr>
<tr>
<td>Manhattan</td>
<td>22.8</td>
<td>1.6</td>
</tr>
<tr>
<td>Queens</td>
<td>108.5</td>
<td>2.3</td>
</tr>
<tr>
<td>Staten Island</td>
<td>58.4</td>
<td>0.5</td>
</tr>
</tbody>
</table>

1. (6 points) Create a scatterplot by plotting each of the boroughs on the axes below (with area as the x-variable and population as the y-variable, as the axis labels indicate). Label each point with the name of the borough and its (x, y) coordinates (as shown for The Bronx).

Solution:

![Scatterplot showing boroughs](chart.png)

2. (2 points) Enter the data into a spreadsheet, and use =correl to calculate the correlation coefficient for the data set:

   correlation coefficient \( r = \)

   **Solution:** \( r \approx 0.469 \)

   This value of the correlation coefficient matches the weak positive correlation shown on the scatterplot (see the next question for reasons why the correlation is positive but weak).
3. (2 points) Your scatterplot and correlation coefficient should indicate a low (or “weak”) positive correlation between area and population. Explain why we might in general expect a positive correlation between area and population, but why the correlation is relatively weak.

(a) We might expect a positive correlation between area and population because...

Solution: ...a greater area of land will have room for more people to live, i.e., will tend to have a greater population;

(b) but the correlation is relatively weak because...

Solution: different areas—in particular, the five boroughs—have different capacities for population density, because they have different land use patterns (e.g., Staten Island has a large number of suburban-style single family homes and large open spaces, whereas Manhattan has many high-rise apartment buildings.)

4. This was not originally on the quiz, but we will go over it in class, and you should review it as preparation for the exam! some solutions for (a)-(c) can be viewed this Google spreadsheet:

https://docs.google.com/spreadsheets/d/1VGJQVJmQY6xnTuW5EwQTcCJU55bWo9ajUtt2Y0i8Fdpce/edit?usp=sharing

(a) Draw in the linear regression trendline on your scatterplot.

(b) Find the linear regression parameters using =slope and =intercept:

• \( y \)-intercept \( \alpha = \)
• slope \( \beta = \)
  so the equation of the regression line is \( \hat{y} = \)

(c) Use the linear model to find the predicted population of a hypothetical “6th borough” which has area 100 square miles? Do the calculation in your spreadsheet, but also write out the calculation below. Then plot this point on the graph, and label it with its \((x, y)\) coordinates.

\[ \hat{y} = \]

(d) Why should area be plotted on the \(x\)-axis and population on the \(y\)-axis when setting up the scatterplot and performing linear regression (i.e., why should area be the independent variable, and population the dependent variable)?