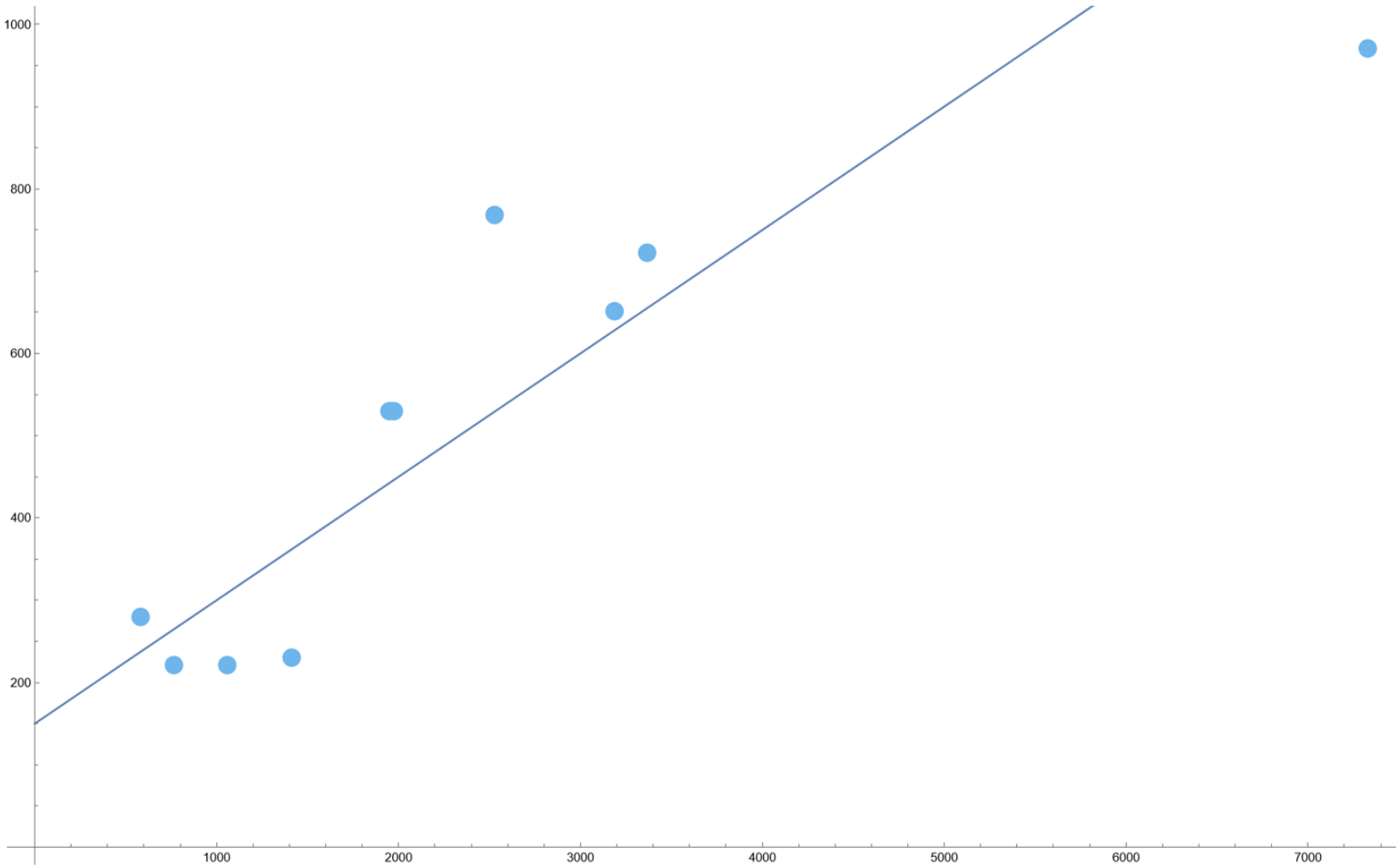
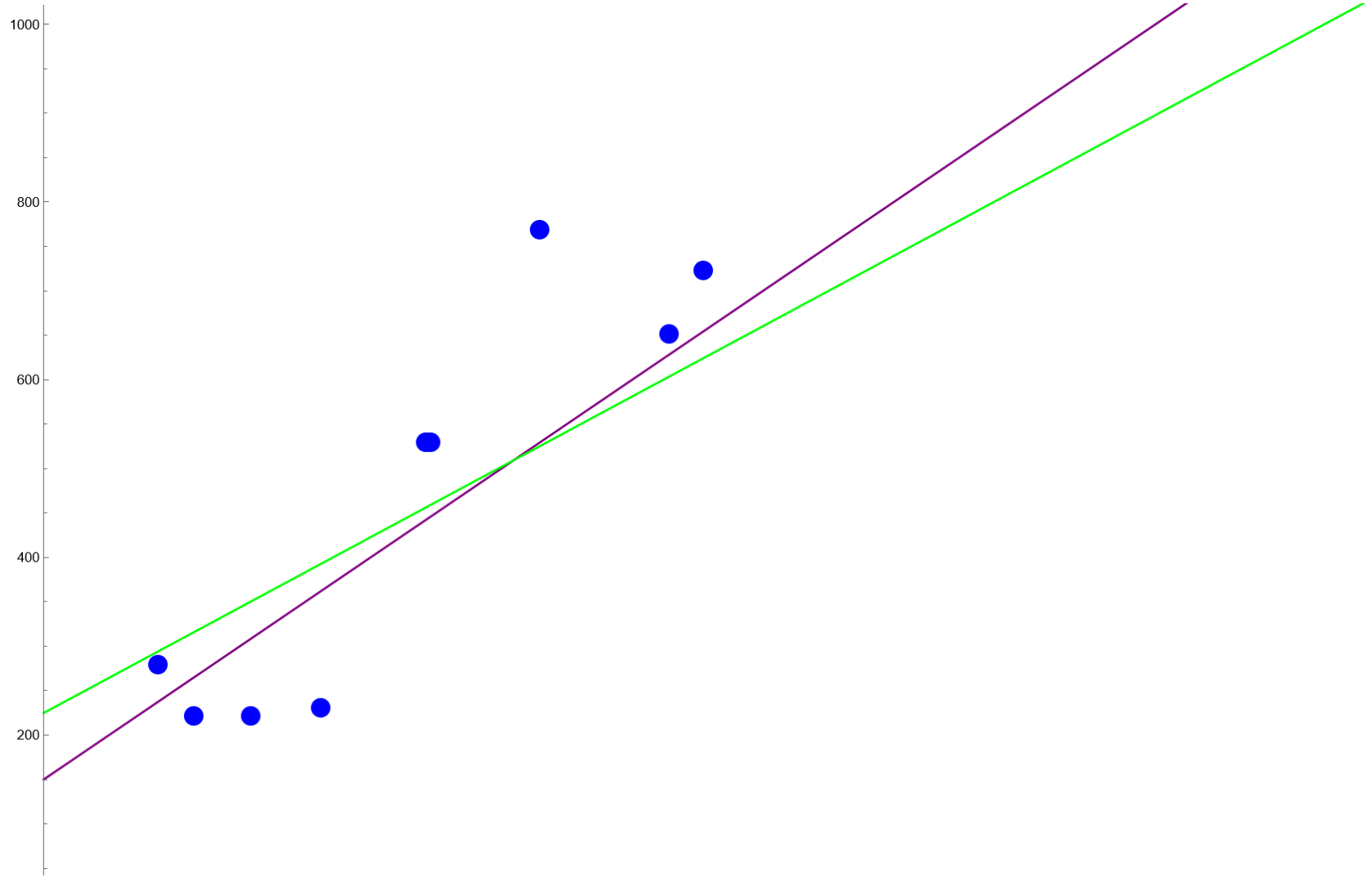
STA Project – Part II

1. Using the graph from STA 225 Project Part I. Copy and paste it below. Use a straight edge to approximate a line of best fit to the data. See page 553 in our textbook for info.


1000
800
600
400
200
1000
2000
3000
4000
5000
6000
7000

This is my guess in purple and the line of best fit in green



2. On a scale of 0 to 1, estimate how well you feel the line fits the data. Use the scale, 0 = *no fit* and 1 = *a perfect fit*. Discuss how you arrived at your estimation and what factors or reasons made you choose the value?

I think 0.83. I weighted the left data points more heavily than the Honolulu data point. The data is close to a linear fit but not on an almost perfect line because of deviations.

3. Find the equation based ***on your best fit line***. HINT: To find the estimated equation, pick two points on (or closest to the line).

   a. The two points chosen are: (___0____, ___150____) and (_1000_____, ___300____)

   b. Calculate the slope between the two points.

   0.15

c. Plug the points into the point-slope formula $y - y_1 = m(x - x_1)$ and solve your equation for y. If you need assistance with this, I've included the links to two Khan Academy videos into the folder. **Indicate the equation below.**

y-300=0.15(x-1000)

4. Now, let's calculate the least-squares line ***based on your data***. Remember, the data uses the ten cities from Part I (Miami, San Diego, NYC, Chicago, Seattle, SLC, Boston, Honolulu, Denver, and the city you chose from your initial discussion board posting.

Use this chart below or replicate the chart in Excel. If you use Excel, you must attach the file when you submit the assignment. **You are being graded on mastering this process, which includes using the formulas as presented in our book with your data.** Do not use Excel functions, other than Sum. Use Guided Example 4 (p. 555) as a guide.

x	y	x^2	xy	y^2
1416	230	2005056	325680	52900
3191	651	10182481	2077341	423801
768	221	589824	169728	48841
586	279	343396	163494	77841
3367	722	11336689	2430974	521284
2531	768	6405961	1943808	589824
1060	221	1123600	234260	48841
7331	970	53743561	7111070	940900
1951	529	3806401	1032079	279841
1976	529	3904576	1045304	279841
$\sum x = 24177$	$\sum y = 5120$	$\sum x^2 = 93441545$	$\sum xy = 16533738$	$\sum y^2 = 3263914$

It is important that you **show your work** for parts a – e.

a. $\bar{x} = \frac{\Sigma x}{n} = 24177/10 = 2417.7$

b. $\bar{y} = \frac{\Sigma y}{n} = 7120/10 = 512$

c. $b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = (10*16533738\text{-}24177*7120)/(10*110653395\text{-}(24177)\text{^}2) = 0.118756$

d. $a = \frac{\Sigma y}{n} - \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}\frac{\Sigma x}{n} = \frac{5120}{10} - \frac{10*165337380 - 24177*5120}{10*934415450 - 58452739}\left(\frac{24177}{10}\right)$ a=224.88482

e. *Write your equation in the form,* $\hat{y} = a + bx$. $\hat{y} = 224.88482 + 0.1187555x$

5. **Show your work** of how you will determine the Sample Correlation Coefficient, r. Hint: Use Guided Exercise 3 (p. 542) as a guide.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{10*165337380 - 24177*5120}{\sqrt{10*934415450 - 584527329}\sqrt{10*3263914 - (5120)^2}} = 0.876376$$

6. Compare your estimate of *r* from **Step 2** to the value of r that you calculated in **Step 5**. Was your estimation close? Why are they not the same?

I estimated 0.83 and the correct answer was 0.876. I think this is a good guess.

I thought the data was worse than the real answer. I underestimated the value of r and the goodness of fit. It might be a good idea to use adjusted r squared or Akaike Information Criterion to take into account the problem of overfitting the data for a more in-depth analysis.

7. Now, you would like to choose ***one additional city***, not previously used. Use the internet to find the distance, flight info, and amount. Remember, you can't choose: Miami, San Diego, New York City, Chicago, Seattle, Salt Lake City, Boston, Honolulu, Denver, or your chosen city picked on the discussion board.

New Destination	**Departing Flight Numbers (list all flight segments)**	**Distance (round to nearest mile) kilometer**	**Amount (Round to nearest dollar)**
Saint Louis and St Louis Lambert International Airport	NK 1217 NK 1131	691 km	270

8. Plug the Distance into the equation found in **Step 4e**. Determine $\hat{y}$. **Show your work.**

y(691)=224.885 + 0.118756 *691=306.945

the difference between the predicted price and the real world price is 36.9449.

9. Was the value of $\hat{y}$, that you found in **Step 8** similar to the **Amount** listed in the table on **Step 7**? Why or Why not?

It was kind of far away and kind of close. I was hoping for more accuracy to around 20 or less dollars like 15 dollars or even 10 dollars or less.

Self-Check:

- Have you shown your work for questions: 3, 4a – 4e, 5, and 8?
- Have you given a well-thought-out answer for questions: 2, 6, and 9?
- Submit the:
    - STA 225 Project Part II Word document
    - Scatterplot with best-fit line
    - Any calculations that you did by hand
    - Any Excel worksheets which **show you using the formulas as found in 9.1 and 9.2**.