**1. Project Overview** [Maximum 100 words] What data source(s) did you use and what technique(s) did you use analyze/process them? What did you hope to learn/create?

In our project, we wanted to find out more about turtles. We wanted to explore what characteristics (physical, geographical, or behavioral) were most crucial to the turtle. We also wanted to find out what types of things should come to mind when people think of turtles. We used the page about turtles from Wikipedia as our data source and used word frequencies and summary statistics to try and answer our questions.

**2. Implementation** [~2-3 paragraphs] Describe your implementation at a system architecture level. You should NOT walk through your code line by line, or explain every function (we can get that from your docstrings). Instead, talk about the major components, algorithms, data structures and how they fit together. You should also discuss at least one design decision where you had to choose between multiple alternatives, and explain why you made the choice you did.

One important feature of our implementation was the way we imported and had our program read our data source of the Wikipedia page. We imported the file through the Wikipedia package and chose to print the content (text) of the turtle specific Wikipedia page in another text file. We then had to save that text file under the Windows section of our PC because the file wasn’t recognizing the path of our source when saved within our assignment folder because there may have been a space or odd character in the name of our Assignment folder. This strategy was learned in class when the same error occurred. Once the file was finally recognized and being read by our program we were able to start testing out different analysis functions. Luckily for us, the Python Wikipedia package, when called upon, already gave us an organized text-only version of the Wikipedia “Turtle” page. This allowed us to easily parse out words without having to write too much code to fix the structure or change the data.

One design decision we had to make was having to do with our possible sources. At first, we contemplated using blogs, scientific websites, or other turtle articles. We decided not to use blogs about turtles because we felt that they might be too biased based off of the purpose for righting their article and the extent to which they like turtles. We thought scientific website wouldn’t be as helpful because their sites are based off of multiple different pages and reliant upon many graphs which would have been difficult to parse through with the skills we have learned in regards to this project. We determined that Wikipedia would be the least biased source and would provide the most efficient platform to parse through data.

**3. Results** [~2-3 paragraphs + figures/examples] Present what you accomplished:

* If you did some text analysis, what interesting things did you find? Graphs or other visualizations may be very useful here for showing your results.
* If you created a program that does something interesting (e.g. a Markov text synthesizer), be sure to provide a few interesting examples of the program's output.

By using word frequency and summary statistics, we were able to come up with the most crucial characteristics to the turtle specie’s story that someone who doesn’t know much about turtles can use to learn more about them within minutes. Here are some interesting deductions that we came up with, simply by analyzing word frequency in a file with 5,333 words:

* “tortoise” or “tortoises” used 32 times compared to “turtle” and “turtles” used 216 times suggests that even though many people think that these two words can be used interchangeably, they most likely don’t mean the same thing and that a tortoise might be a variation of a turtle.
* “family” was used 50 times showing that there are definitely many different types of turtles that exist or have existed because with family being used to make sure the reader knows which family of turtle is being discussed.
* The words “freshwater” was used 14 times, “sea” was used 33 times, “aquatic” was used 7 times shows that a large proportion of the types of turtles that exist, can typically be found in the water.
* The word “extinct’ was used 9 times, “endangered” was used 6 times, and “evolution” was used 8 times showing that turtles have existed for a very very long time and have continuously evolved over that period. This also shows another reason why turtles have so many different types. Along that period many have become extinct and many are currently still considered endangered.
* The word “reptile” was used 7 times suggesting that turtles must be reptiles.
* The word “florida” was used 6 times and more than any other location word which suggests that turtles might have a very high population in the state of Florida.

**4. Reflection** [~1 paragraph] From a process point of view, what went well? What could you improve? Other possible reflection topics: Was your project appropriately scoped? Did you have a good plan for unit testing? How will you use what you learned going forward? What do you wish you knew before you started that would have helped you succeed?

Also discuss your team process in your reflection. How did you plan to divide the work (e.g. split by task, always pair program together, etc.) and how did it actually happen? Were there any issues that arose while working together, and how did you address them? What would you do differently next time?

Overall, we think our ability to relate word frequency to describing the main characteristics of turtles was pretty strong. However, we feel like we could have strengthened that by omitting words like “the”, “and, and “a” in our word frequency generations because they don’t help our argument or findings and just make our results more convoluted. We also feel like we could have used another Wikipedia page about a similar animal to compare the different characteristics between the two species as well as what they have in common. We will use our analysis and new word frequency understanding and apply it to our term project in a way that will help optimize our linear regression for our bid calculator. In terms of team work, we worked really well together. We used to approach of working together at all times where one team member typed the code while the other was by their side helping. The roles were switched for writing the paper as the other partner that before helped, now typed the paper while the other helped them.