**TP1 Project Proposal**

Cindy Luo

4.16.22

**Project name: skim(Data)**

**Project Description:**

This project aims to write an algorithm that allows the user to import a dataset and generate different charts and data visualizations through an interactive dashboard with just a few clicks. It will be similar to an ultra-lite version of excel but will be really handy. Using points & clicks to generate the graphs requires no prior experience with data from the user. The design of the algorithm will make it very accessible to the users.

It will quickly provide users with an overview of a data frame. It will also suggest interesting directions & questions by visualizing the potentials correlations between the variables through clear visualizations.

**Competitive Analysis**

Most data visualization projects using Python use libraries such as pandas to process data and matplotlib to generate different types of graphs and charts. Though it is beneficial to learn how to use these libraries, it is also a great exercise to start everything from scratch. By searching for how to work with data in python without using the libraries, I found one of the blogs online that suggested viewing the data frame as a 2d List. It was a great starting point for my project. In addition, from the term project gallery, I found another student that completed a data visualization project. They embedded some basic statistical analysis techniques in their visualization project, and I thought adding this feature would improve the complexity of my project, as well. The main difference between my project and others is that I want to focus on data storytelling. I want my visualizations to include not just simple x and y variables that show some type of relationship but a relatively meaningful visualization that could suggest exciting directions for follow-up analysis.

**Structural Plan**

Class Variable & Class Dataframe:

I will create two classes to store functions that are related to data manipulations on the back end. These functions will allow the program to process & clean the data first before visualizing it, and it will allow generalization to different types of datasets.

Interactive board:

These sets of codes will be mainly used for user experience and the “points & clicks” part of the program. Different modes will be created.

Visualization:

1. Visualizing the data in a table:

Once imported a dataset, it is always useful to visualize the data in a table-like format (like the ones in excel)

1. Data visualization:
   1. Bar chart
   2. Pie chart
   3. Scatter plot
   4. Display of the descriptive statistics and the result of the statistical analysis
2. Data analysis
   1. Mathematical operations to perform linear regression
   2. Mathematical operations to perform multiple regression

**Algorithmic Plan**

The interactive part and the data analysis part will be the algorithmically most complex parts. I haven’t started working on those parts yet…

**Timeline Plan**

* Generating different types of charts (pie, bar, line, etc)
  + Bar chart (Saturday Night, April 16)
  + Pie chart (Sunday, April 17)
* Analyzer (linear regression, other regression algorithm(s))
  + Linear (Monday, 4/18)
  + Multi-regression (Wednesday-Thursday)
* Thursday and Friday (4/21 & 4/22)
  + Design the interactive board

**Version Control Plan**

I uploaded everything into my Google drive, and I will update what I have once per day.

Graphical user interface, application, Teams

Description automatically generated

**Module List:**

No modules were used except cmu\_112\_graphics.py.

**Changes & Updates TP2:**

See the new storyboards for details, added a lot more interactive features & advanced graph features.

~~Main changes: Multiple linear regression to polynomial regression~~

**Changes & Updates TP3:**

Did not get to complete all the features demonstrated in the second version of the storyboard.

Main changes since TP2:

- Achieved MVP by embedding multiple linear regression in my analysis section & displaying the result

- Added a short description of the project in the “infoMode"

- Allowed the user to import a data set from a CSV file

- Created an interactive board & implemented the interactive board for all features

- Double-checked everything & corrected any obvious bugs

- Went through feedback collected from my peers and made improvements based on the comments:

The feedback I received & Improvements I made:

- Try to prevent error messages by helping the user determine the correct data type (categorical vs. numerical) for the graphs or consider creating different classes for data

[Fixed it by disallowing the inappropriate variables to be chosen in the first place]

- Case-sensitive input & a lot of typing

[Fixed it by replacing user input with buttons]

- Default Database to prevent importing database multiple times when there is an error message

[Fixed it by creating a shortcut command to load demo data]

- X-axis labels overlap with each other when "Date" is the x-variable

[Fixed it by adjusting the angle of the text]