Tianyi Kong
Professor Brooke Luetgert
Data Visualization for Humanities
Project Plan

1) What data source will you use? #observations? time frame or other indications of data range covered.

I choose the dataset that comes from Kaggle, focusing on e-commerce sales on Xiaohongshu, a Chinese social commerce platform. It contains 29,452 records, covering user purchasing behavior, engagement, and demographics.

Key data fields include:

- Revenue: Purchase amount per order.
- 3rd_party_stores: Whether the user has purchased from third-party stores.
- Gender: 1 for male, 0 for female (blank if unknown).
- Engaged_last_30: Whether the user has engaged in activities (discussions, photo sharing) in the past 30 days.
- Lifecycle: User registration time category (6 months, 1 year, 2 years).
- days_since_last_order: Number of days since the user last placed an order.
- previous_order_amount: The user's cumulative purchase history.

2) What will the visualization convey?

This dataset is suitable for exploring user purchasing behavior on Xiaohongshu and applying data visualization and machine learning techniques to predict future sales trends. For example, I can analyze spending patterns over different lifecycle stages to show user purchasing trends. I can also compare revenue distribution between male and female users to see gender-based purchase differences.

3) What two methods are you combining in your project?

I am going to use statistical analysis and visualization tequegiques to explore and interpret the dataset. For example, I will apply mean, median and standard deviation to understand general trends in user purchasing behavior. For predictive insights, I might use regression analysis, like linear regression, to model user spending behavior based on past purchase history.

Moreover, for data visualization techniques, I may want to create bar charts, line graphs, and scatter plots to visually explore trends and relationships within the data. I will also think about creating histograms and box plots to analyze the distribution of revenue and previous order amounts.

4) Do you have any concerns for your planned project?

One of the main concerns for my project is dealing with missing values in key variables such as gender, age, and engaged_last_30. Missing data can impact the accuracy and reliability of the analysis, so I will need to carefully decide how to handle these gaps.

Beyond missing values, I also need to think about other quality concerns, like outliers or inconsistent data entries, that could mess with the analysis. Making sure the data is solid will be a very important step before diving into any more analysis.