1. ***Name of the project and team members***

   Analyzing the predictive factors of Top 100 K-pop group album sales
   Feier Su

2. ***What problem are you trying to solve?***

   Analyzing the correlations between predictive factors and physical album sales

3. ***How will you collect data and from where?***

   Dependent Variable: The top 100 artists in terms of total physical album sales (groups only)
   https://soridata.com/sales/?rank=sales

   Independent Variables:

   1. **Basic Info**
      - Company label: e.g. HYBE, SM, JYP, YG, etc. (via Wikipedia
      - Gender: boy or girl group (via Wikipedia
   2. **Streaming Performance**
      - Digital scores https://soridata.com/kpop_streaming.html?rank=score&gto=1&gendero=0
      - YouTube streams https://soridata.com/en/mvs.html?options=101&tags=dhcmoravxpnezjsl&cols=100110100&ord=1D
   3. **Reputation & industry recognition**
      - Number of Korean music show awards https://soridata.com/awards_by_artist.html?showrank=true&gto=1&gendero=0
      - Number of annual awards https://soridata.com/en/yawards.html?gto=1&gendero=0
   4. **Online engagement/ Fandom size**
      - Number of Instagram followers https://www.kpop-radar.com/instagram?type=2&date=1&gender=1
      - Number of Ytb followers https://www.kpop-radar.com/youtube?type=2&date=1&gender=1
      - Number of Tiktok followers https://www.kpop-radar.com/tiktok?type=2&date=1&gender=1
      - Spotify followers https://www.kpop-radar.com/spotify?type=2&date=1&gender=1
   5. **Offline engagement**
      - Number of concerts since debut via Wikipedia

4. ***What analysis will you do and what visualizations will you create?***

   Model: Multiple Linear Regression and Random Forest Model

   Visualizations: scatterplots between IVs and DV, Feature Importance Plot, pie charts, etc.