Data Cleaning Final Project Proposal

Group members: **Niken Shah (niken2), Shubhangi Singhal (ss100)**

1. **Purpose of your project**

Describe what you intend to do for the final project

We intend to work on the Google Play Store Dataset and perform data wrangling tasks to prepare it for further Exploratory Data Analytics. We will also be developing a data visualization dashboard to get some insights from the data.

We would be utilizing the following data wrangling tasks:

1. Formatting the data values across columns for standardization
2. Renaming any field names for better audience readability and understanding such as *App and Installs*
3. Formulating process to handle NaN values in the dataset
4. Applying advanced sorting functions to get top applications on our dataset
5. Rectifying any incorrect data types of the columns such as for *Reviews*
6. Perform data value clustering on column *Type*
7. Handling blank and incorrect values

The objective of our EDA would be to identify:
Top applications based on highest ratings received and maximum installations both basis the category of the application, price, and type of the application.

What are the use cases of the project?

Our use case is to create an interactive dashboard with customizable filters for the end users, based on various columns, after we are finished with the data cleaning section. Any user (e.g. business organization) can just make use of our dashboard to derive insights instead of being limited to just reading through the original data.

What is the purpose/significance/contribution of this project?

Our data cleaning project would be useful for organizations which have their applications listed on the Google Play Store. They can look at the reach and performance of their competitors in the respective categories based on metrics like total downloads, average rating, number of reviews and application size. Similarly, someone planning to launch an application can look at potential competitors in the market. The main purpose would be to help users who want to download an Android application for a certain task like travel booking or video calling. They can look at alternatives based on review, rating and price and make an informed decision.

Cite some references or literatures to support your ideas/ hypothesis (if any)

1. https://elitedatascience.com/data-cleaning

2. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7889976

2. Dataset Description

Where did you get the dataset from?

Our dataset can be found on Kaggle. Please use the link to access the same : https://www.kaggle.com/lava18/google-play-store-apps

What is it about?

The dataset contains details about the Android applications available on Google Play Store.

What fields/headers do the dataset have?

The following are the attributes/ fields of the dataset –
App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Version and Android Version.

How many columns/rows
There are 13 columns and about 10,841 rows. Our dataset is a comma separated values (.CSV) file type.

3. Data Cleaning processes

Data cleaning tools you intend to use (if any)

Python and Open-refine tool

Other methods or steps (not necessarily data cleaning) that you will be using

We will use Tableau or PowerBI for data visualization by creating a dashboard for the end user.

4. Project deliverables

Describe the desirable outcomes you wish to see from what you are doing for this project

Exploring different tools for data cleaning and having a clean, structured, and well formatted file for further analysis. We will then use the clean data file to find some insights from the data using visualization. E.g.: Finding the average **rating** of applications based on the **category** column from our data set.

The following would be the scope of our deliverables –
1. Source Dataset
2. Project Proposal and Project Report
3. Cleaned dataset and data visualization file in Tableau (along with a link of the dashboard)