Update-level Variables

1. Variables related to description length
   - **len_word:** total number of words in a given update description, computed by counting the words after tokenization and removal of special characters (e.g., '•', '*', '◉').
   - **len_char:** length of characters in words used in the calculation of len_word.
   - **len_char_ttl:** total length of all characters in an update description.
   - **Para:** number of paragraphs in an update description
2. Variables related to revision content
   - **feature_score**: the percentage of feature-related words in all normalized words of a given update description. Normalization includes steps such as reformatting characters to lower cases, removing stop words, and lemmatization. I created a dictionary of feature-related words based on the dataset.
   - **bug_score**: the percentage of bug-related words in all normalized words of a given update description. I created a dictionary of bug-related words based on the dataset.
   - **system_mention**: 1 if the description mentions operation system update, 0 if not. I identify mentions by searching over a dictionary of system update-related words I created based on the dataset.
   - **privacy_mention**: 1 if the description mentions operation system update, 0 if not. I identify mentions by searching over a dictionary of system update-related words I created based on the dataset.
3. Variables related to version number
   - **update_num_dist**: the first non-zero difference between digits of the current version number and the lagged version number. E.g., 1.2.0 -> 1.2.2, update_num_dist = 2; 1.2.2 -> 1.3.0, update_num_dist = 1
   - **min_max_diff**: the difference between the largest digit place modified and the smallest digit place modified during the sample period. E.g., if a given app has three updates 1.2.0, 1.2.2, and 2.0, then the largest digit place modified is the third digit (1.2.0 -> 1.2.2), and the smallest digit place modified is the first digit (1.2.2 -> 2.0). the min_max_diff is 3-1 = 2.
4. Variables related to major revision measures
   - **update_by_direct_mention**: 1 if the update is identified as major revision based on direct mention, 0 if not. A dictionary of words related to major revision is developed to search for direct mentions.
   - **update_by_num_modified**: 1 if the update is identified as major revision based on changes of version number, 0 if not. For apps with more than three levels of update (max_min_diff <= 2), I classified updates with changes in the smallest digit to be major updates; For apps with less than three levels of update, I classified them based on the distance measure (distance larger than the 75^{th} quantile for the given app). This method accommodates cases where only digits at a certain location change throughout the sample period, e.g., only the third digit changes, 0.0.2 all the way to 0.0.80.

- **update_by_feature_bug**: 1 if the update is identified as major revision based on feature and bug score – feature score larger than 75th quantile for the given app and bug score smaller than 25th quantile for the given app, 0 if not. (Revisions related to core features of the app tend to be major updates whereas revisions that only involves bug fixes may be minor revisions or maintenances.)
- **update_by_para**: 1 if the update is identified as major revision based on the number of paragraphs, 0 if not. If the number of paragraphs exceed the 90th percentile for the given app, the update is classified as major revision. (Rationale: descriptions with multiple paragraphs may involve more aspects of revision.)

App-level Variables

(Please refer to details of the variables in the section above)

1. Variables related to description length
   - **avg_len_word:** average number of words for all descriptions of a given app
   - **avg_len_char:** average length of characters for all descriptions of a given app
   - **avg_len_char_ttl:** average length of characters for all descriptions of a given app
   - **avg_para:** average number of paragraphs for all descriptions of a given app
2. Variables related to number of updates
   - **ttl_update**: total number of updates over the sample period for a given app
   - **avg_monthly_update**: average number of updates monthly over the sample period for a given app
   - **update_by_direct_mention**: total number of major revisions identified by direct mention over the sample period for a given app
   - **update_by_num_modified:** total number of major revisions identified by change of version number over the sample period for a given app
   - **update_by_feature_bug**: total number of major revisions identified by feature and bug score over the sample period for a given app
   - **update_by_para**: total number of major revisions identified by number of paragraphs over the sample period for a given app
3. Variables related to revision content
   - **avg_feature_score**: average feature score for all descriptions of a given app
   - **bug_score**: average feature score for all descriptions of a given app
   - **system_mention**: number of descriptions mentioning OS update during the sample period for a given app
   - **privacy_mention**: number of descriptions mentioning privacy related words during the sample period for a given app