HiCounselor

Playstore Apps analysis & Visualization

About the project:

In this project, you will be working on a real-world dataset of the google play store, one of the most used applications for downloading android apps. This project aims on cleaning the dataset, analyze the given dataset, and mining informational quality insights. This project also involves visualizing the data to better and easily understand trends and different categories.

Project Description:

This project will help you understand how a real-world database is analyzed using SQL, how to get maximum available insights from the dataset, pre-process the data using python for a better upcoming performance, how a structured query language helps us retrieve useful information from the database, and visualize the data with the power bi tool.

The Project will consist of 2 modules:

Module 1: Pre-processing, Analyzing data using Python and SQL.
Module 2: Visualizing data using Power bi.

Prerequisites for the Project (mandatory):

1. SQL (MYSQL)
2. POWER BI
3. Excel
4. Python

Module 1

Pre-processing, Analyzing data using SQL

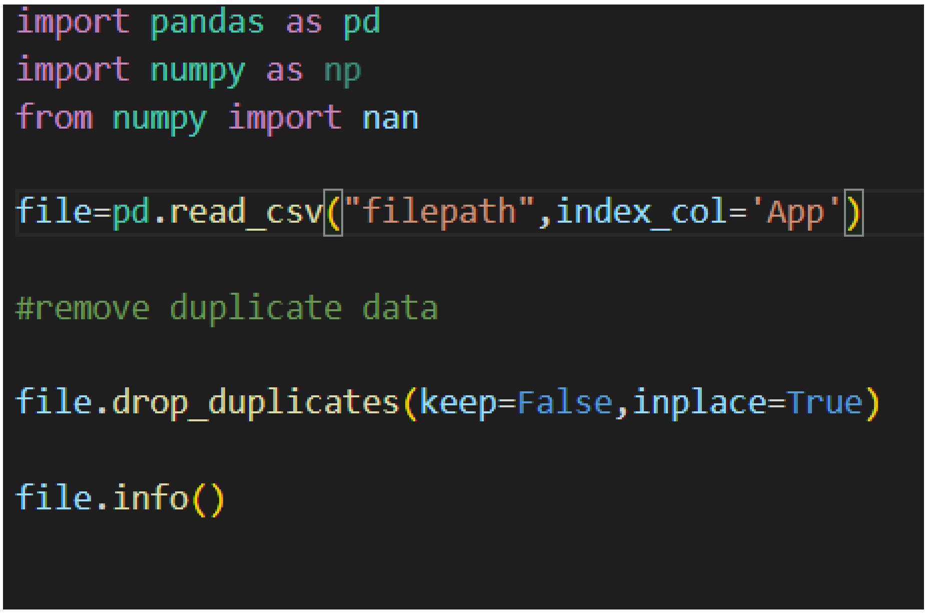
In this module, you will query the dataset using structured query language to gain insights from the database. The problem statements to be solved will be provided to you and you need to provide the solution for the same using your logic. Different concepts of SQL will be used in this process such as aggregating the data, grouping the data, ordering the data, etc. Module 1 consists of subtasks which are as follows

Task 1: Pre-processing the data

Data Pre-processing is one of the important steps in data analytics because data that is not processed can lead to different unwanted results when the data will be used for further applications. This task includes sub-tasks such as handling null values, deletion or transformation of irrelevant values, datatype transformation, removing duplicates, etc. The tasks to be performed for cleaning the data set are given below:

Here is the link to the data set for the module : Click Here

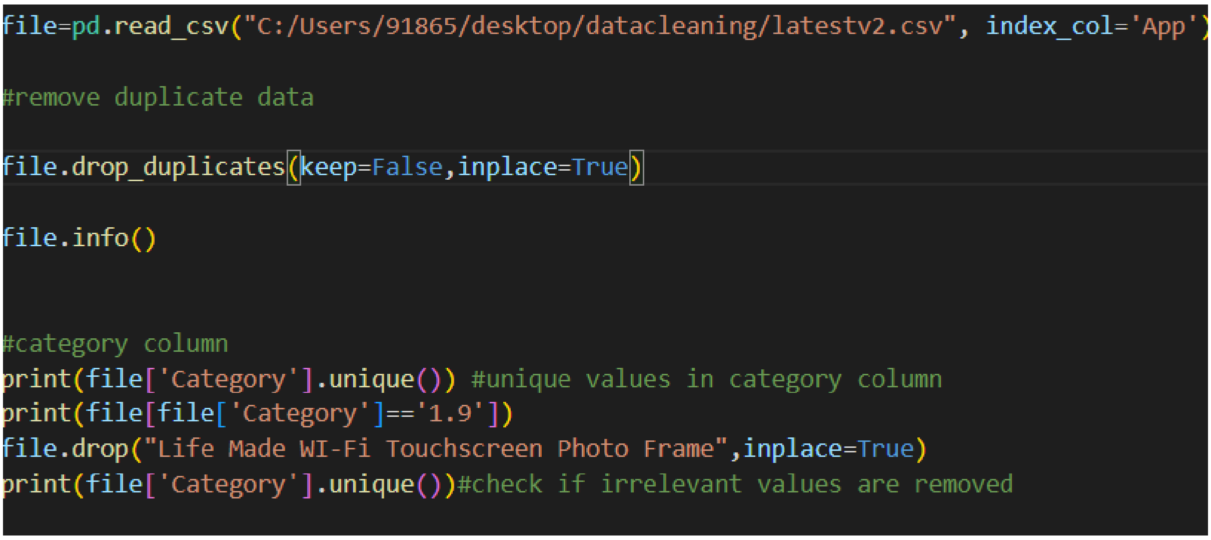
Subtask 1: Removing duplicate rows:



```
import pandas as pd
import numpy as np
from numpy import nan

file=pd.read_csv("filepath",index_col='App')

#remove duplicate data

file.drop_duplicates(keep=False,inplace=True)

file.info()
```

Subtask2: Remove irrelevant values from each column if any:

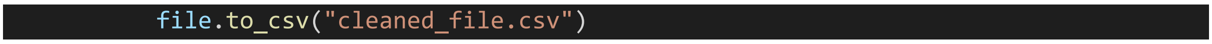
Use the .unique function in python provided by the NumPy library to get a list of all the unique values present in that particular column, if a unique value does not comply with the data type of the column or does not justify the characteristic of the attribute then remove all the rows having that unique value.



```
file=pd.read_csv("C:/Users/91865/desktop/datacleaning/latestv2.csv", index_col='App')

#remove duplicate data

file.drop_duplicates(keep=False,inplace=True)

file.info()


#category column
print(file['Category'].unique()) #unique values in category column
print(file[file['Category']=='1.9'])
file.drop("Life Made WI-Fi Touchscreen Photo Frame",inplace=True)
print(file['Category'].unique())#check if irrelevant values are removed
```

In the same way, perform the above operations for all the columns, and check if all the unique values are of the same data type.

Subtask 3: Export the cleaned dataset as a .csv file and prepare the data using excel:



```
        file.to_csv("cleaned_file.csv")
```

Open the exported file in excel and check for the null values in each column using the find and replace tool in excel.

Note: keep the find what section empty to search for null values

Hint:

1. As the rating can be a float value change the null values of the rating column to 0
2. Delete the row with a null value for columns with datatype text or string

3. For column current ver if there is any null value, change it to NaN
4. In app reviews dataset, remove all the rows with NaN reviews as there is no translated review present to analyze, hence these kind of data is of no use.

Subtask 4: Encoding data into suitable format

The dataset we used in this project was populated using a scraping technique, the encoding method used may be different, we prefer UTF-8 encoding as it is majorly used in all the database servers. To do so perform the below tasks:

1. In the menu bar of excel select option data. Then select from text/CSV and choose the file in which u made changes using excel
2. The default file origin will be UTF -8 keep all the field as it is and select load
3.Save the file as CSV(comma delimited) file.
4. Import the file into your SQL database
(Analyze the dataset with respect to your database system and then change it, if necessary).

Go through both cleaned datasets for a perfect understanding:

Understand the following requirements and query the dataset for displaying the required solution:

1. Which apps have the highest rating in the given available dataset?

2. What are the number of installs and reviews for the above apps? Return the apps with the highest reviews to the top.

3. Which app has the highest number of reviews? Also, mention the number of reviews and category of the app

4. What is the total amount of revenue generated by the google play store by hosting apps? (Whenever a user buys apps from the google play store, the amount is considered in the revenue)

5. Which Category of google play store apps has the highest number of installs? also, find out the total number of installs for that particular category.

6. Which Genre has the most number of published apps?

7. Provide the list of all games ordered in such a way that the game that has the highest number of installs is displayed on the top
(to avoid duplicate results use distinct)

8. Provide the list of apps that can work on android version 4.0.3 and UP.

9. How many apps from the given data set are free? Also, provide the number of paid apps.

10. Which is the best dating app? (Best dating app is the one having the highest number of Reviews)
11. Get the number of reviews having positive sentiment and number of reviews having negative sentiment for the app **10 best foods for you** and compare them.

12. Which comments of **ASUS SuperNote** have sentiment polarity and sentiment subjectivity both as 1?

13. Get all the neutral sentiment reviews for the app **Abs Training-Burn belly fat**

14. Extract all negative sentiment reviews for **Adobe Acrobat Reader** with their sentiment polarity and sentiment subjectivity

Upon completing the module, please submit your ZIP files in the google form here.

Module 2

Visualizing data using Power BI

In this module, you will be using both datasets for visualizing what the numbers in and behind the data want to convey. You have to create a dashboard for the same using different statistical graphs and diagrams for a visual understanding and analysis. Given below are the requirements, create a dashboard consisting of mentioned visuals.

1. Design a dashboard of your choice and imagination with attractive wallpapers and designs of visuals

2. The dashboard must consist of basic power bi visuals like stacked bar charts, cards, line charts, pie charts, etc

3. You can use any visual capable of representing the given dataset.

4. Columns having numeric data types are to be majorly used in visuals

5. Include both datasets.

****Upon completing the module, please submit your files in the google form here.***

.

Weekly Module Progress Submission Process:

Each module is to be completed in one week. Submissions will be on weekly basis. Module 1 submission is to be done in the first week and module 2 submission in the next consecutive week.

Guidelines for submissions are mentioned below:

1. A google form will be circulated for module files submissions.
2. For module 1 you have to submit 4 files as mentioned below:
   1. A Text file consisting of the solutions to each SQL requirement.
   2. .py file used for cleaning the dataset
   3. Cleaned .csv file of playstore_apps dataset
   4. Cleaned .csv file of playstore_reviews dataset

**create a zip file of all the 4 files and then make the submission.*

3. A Youtube video link that consists of a screen-recorded solution of the query. (Run the query and display its solution).

For module 2 you have to submit :

1. A power bi file (.pbix)
2. An Youtube video link that consists of a screen-recorded solution of the designed dashboard.

Evaluation Criteria:

1. The project submission guidelines and prerequisites are mandatory and non-negotiable.
2. The submissions that are completed will be considered for evaluation.

Session Schedule & Timelines:

Day	Date	Time	Session
Wednesday	**4th January**	**6:00 PM PST**	**Project Kickoff Call/ Introduction to Module 1**
Saturday	**10th January**	**6:00 PM PST**	**Doubt Solving Session for Module 1**
Wednesday	**11th January**	**6:00 PM PST**	**Module 1 Submission**
Wednesday	**11th January**	**6:00 PM PST**	**Introduction to Module 2**
Sunday	**15th January**	**6:00 PM PST**	**Doubt Solving Session for Module 2**
Thursday	**19th January**	**6:00 PM PST**	**Module 2 submission**

**Any changes in the dates and time of the sessions will be informed in advance.*

Important Links related to the Live Project:

1. Data Live Project Kick Off/ Introduction Call 4th July ‘23 6:00PM PST- click here to register and join
2. Module Submission google form: click here
3. Live Project Doubts and queries: click here