Course 2: Python Project 04

Week 04: Graded Project

PART I:

EDA & Data Preprocessing on Google App Store Rating Dataset.

Domain: Mobile device apps

Context:

The Play Store apps data has enormous potential to drive app-making businesses to success. However, many apps are being developed every single day and only a few of them become profitable. It is important for developers to be able to predict the success of their app and incorporate features which makes an app successful. Before any such predictive-study can be done, it is necessary to do EDA and data-preprocessing on the apps data available for google app store applications. From the collected apps data and user ratings from the app stores, let's try to extract insightful information.

Objective:

The Goal is to explore the data and pre-process it for future use in any predictive analytics study.

Data set Information:

Web scraped data of 10k Play Store apps for analyzing the Android market. Each app (row) has values for category, rating, size, and more.

Attribute Information:

Slno.	Attribute	Description
1.	App	Application name
2.	Category	Category the app belongs to.
3.	Rating	Overall user rating of the app
4.	Size	Size of the app
5.	Installs	Number of user reviews for the app
6.	Type	Paid or Free
7.	Price	Price of the app
8.	Content Rating	Age group the app is targeted at - children/Mature 21+ /Adult

9.	Genres	An app can belong to multiple genres (apart from its main category). For eg. a musical family game will belong to Music, Game, Family genres.
10.	Last Updated	Date when the app was last updated on play store.
11.	Current Ver	Current version of the app available on play store.
12.	Android Ver	Min required Android Version.

Questions:-

1. Import required libraries and read the dataset.

2. Check the first few samples, shape, info of the data and try to familiarize yourself with different features.

3. Check summary statistics of the dataset. List out the columns that need to be worked upon for model building.

4. Check if there are any duplicate records in the dataset? if any drop them.

5. Check the unique categories of the column 'Category', Is there any invalid category? If yes, drop them.

6. Check if there are missing values present in the column Rating, If any? drop them and and create a new column as 'Rating_category' by converting ratings to high and low categories(>3.5 is high rest low)

7. Check the distribution of the newly created column 'Rating_category' and comment on the distribution. comment pending

8. Convert the column "Reviews'' to numeric data type and check the presence of outliers in the column and handle the outliers using a transformation approach.(Hint: Use log transformation)

9. The column 'Size' contains alphanumeric values, treat the non numeric data and convert the column into suitable data type. (hint: Replace M with 1 million and K with 1 thousand, and drop the entries where size='Varies with device')

10. Check the column 'Installs', treat the unwanted characters and convert the column into a suitable data type.

11. Check the column 'Price' , remove the unwanted characters and convert the column into a suitable data type.

12. Drop the columns which you think redundant for the analysis.(suggestion: drop column 'rating', since we created a new feature from it (i.e. rating_category) and the columns 'App', 'Rating' ,'Genres','Last Updated', 'Current Ver','Android Ver' columns since which are redundant for our analysis)

13. Encode the categorical columns.

14. Segregate the target and independent features (Hint: Use Rating_category as the target)

15. Split the dataset into train and test.

16. Standardize the data, so that the values are within a particular range.

PART II:

Data Visualization on Honey Production dataset using seaborn and matplotlib libraries.

Domain: Food and agriculture

Context:

In 2006, a global concern was raised over the rapid decline in the honeybee population, an integral component to American honey agriculture. Large numbers of hives were lost to “Colony-Collapse-Disorder”, a phenomenon of disappearing “worker-bees” causing the remaining “hive-colony” to collapse. Speculation around the cause of this disorder points to hive-diseases and pesticides harming the pollinators, though no overall consensus has been reached. Twelve years later, some industries are observing recovery but the American honey industry is still largely struggling. The U.S. used to locally produce over half the honey it consumes per year. Now, honey mostly comes from overseas, with 350 of the 400 million pounds of honey consumed every year originating from imports. This dataset provides insight into honey production supply and demand in America by state from 1998 to 2012.

Objective:

The Goal is to use Python visualization libraries such as seaborn and matplotlib to investigate the data and get some useful conclusions.

Attribute Information:

Slno.	Attribute	Description
1.	numcol	Number of honey producing colonies.
2.	yield percol	Honey yield per colony. (Unit is pounds)
3.	total prod	Total production (numcol x yieldpercol). (Unit is pounds)
4.	price per lb	Refers to average price per pound based on expanded sales. Unit is dollars.
5.	prodvalue	Value of production (total prod x priceperlb). Unit is dollars.
6.	Stocks	Refers to stocks held by producers. Unit is pounds
7.	Year	Calendar year.
8.	State	Different states' names.

Questions:-

1. Import required libraries and read the dataset.

2. Check the first few samples, shape, info of the data and try to familiarize yourself with different features.

3. Display the percentage distribution of the data in each year using the pie chart.

4. Plot and Understand the distribution of the variable "price per lb" using displot, and write your findings.

5. Plot and understand the relationship between the variables 'numcol' and 'prodval' through scatterplot, and write your findings.

6. Plot and understand the relationship between categorical variable 'year' and a numerical variable 'prodvalue' through boxplot, and write your findings.

7. Visualize and understand the relationship between the multiple pairs of variables throughout different years using pairplot and add your inferences. (use columns 'numcol', 'yield percol', 'total prod', 'prodvalue','year')

8. Display the correlation values using a plot and add your inferences. (use columns 'numcol', 'yield percol', 'total prod', 'stocks', 'price per lb', 'prodvalue')

Submission:

- Please submit the solution files in .html and .ipynb format on Olympus.
- Add necessary comments wherever required.