Note: If you face any difficulty or you want to discuss the query to the mentor feel free to contact the Mentor over LinkedIn - https://www.linkedin.com/in/shivavashishtha/

App Insights Unlocked: A Data Analytics Challenge

Background:

Imagine you are a data analyst working for a tech company that specializes in developing mobile applications. Your company has recently acquired a dataset from the Google Play Store, which contains information about various apps, including their ratings, reviews, sizes, and more. The company is keen to analyze this data to gain insights that can help improve their app development strategies, enhance user experience, and increase app downloads and revenues.

Problem Statement:

Your task is to analyze this dataset to uncover patterns and trends that can guide the company's decision-making process. The analysis should focus on understanding what factors contribute to high app ratings, identifying the most popular app categories, and exploring the impact of app size and price on user reviews and installs.

Stakeholders:

Internal Stakeholders:

- App Developers
- Product Managers
- Marketing Team
- Senior Management

External Stakeholders:

- App Users
- Advertisers
- Partners (e.g., other tech companies or platforms)

Problem Definition:

Your main objective is to analyze the provided dataset to address the following questions:

- What are the key factors that contribute to an app's success on the Google Play Store?
- How can the company leverage these factors to improve its app offerings?
- What insights can be derived about user preferences and behaviors?

Note: If you face any difficulty or you want to discuss the query to the mentor feel free to contact the Mentor over LinkedIn - https://www.linkedin.com/in/shivavashishtha/

Dataset

https://www.kaggle.com/datasets/lava18/google-play-store-apps?select=googleplaystore.csv

Data Requirements:

The data provided includes the following columns:

- `App`: Name of the app
- `Category`: Category of the app
- `Rating`: Average user rating of the app
- `Reviews`: Number of user reviews
- `Size`: Size of the app
- `Installs`: Number of installs
- `Type`: Whether the app is free or paid
- `Price`: Price of the app
- `Content Rating`: Age group for which the app is appropriate
- `Genres`: Genres of the app
- `Last Updated`: Date when the app was last updated
- `Current Ver`: Current version of the app
- `Android Ver`: Minimum Android version required

Metric Development:

Develop metrics to measure app success, such as average rating, total installs, and number of reviews. Analyze these metrics about other variables in the dataset to uncover actionable insights.

Insights & Actions:

Perform targeted analysis to identify high-performing categories, the impact of app size on user ratings, and the relationship between app price and user reviews.

Communication:

Prepare a final report summarizing your findings, including visualizations and recommendations for the company's app development strategy.

Note: If you face any difficulty or you want to discuss the query to the mentor feel free to contact the Mentor over LinkedIn - https://www.linkedin.com/in/shivavashishtha/

Data Dictionary:

Column Name	Description
App	Name of the app
Category	Category of the app
Rating	Average user rating of the app (0-5 scale)
Reviews	Number of user reviews
Size	Size of the app (in MB)
Installs	Number of installs (e.g., 1,000+, 10,000+)
Type	Type of the app (Free or Paid)
Price	Price of the app (if paid)
Content Rating	Age group for which the app is appropriate
Genres	Genres of the app
Last Updated	Date when the app was last updated
Current Ver	Current version of the app
Android Ver	Minimum Android version required

Data Cleaning and Preprocessing:

1. **Handle Missing Values:** Check for and handle any missing values in the dataset.
2. **Convert Data Types:** Ensure data types are appropriate for analysis (e.g., convert `Installs` and `Price` to numerical values).
3. **Remove Duplicates:** Check for and remove any duplicate entries.
4. **Standardize Text:** Ensure consistency in text fields (e.g., standardize `Category` and `Genres`).
5. **Date Formatting:** Convert the `Last Updated` column to a datetime format.

Basic-Level Questions:

1. What is the average rating of apps in the dataset?

- **Hint:** Use `pandas` to calculate the mean of the `Rating` column.
- **How it helps:** Understanding the general user satisfaction level with apps.

- **Business Impact:** Identifying the average satisfaction can help set a benchmark for new apps and improve quality standards.

2. How many unique categories of apps are there?

- **Hint:** Use `pandas` to find the unique values in the `Category` column.
- **How it helps:** Identifying the diversity of app categories available.
- **Business Impact:** Understanding the variety of app categories can help in identifying market opportunities and potential areas for new app development.

3. What is the distribution of app sizes?

- **Hint:** Use `matplotlib` or `seaborn` to create a histogram of the `Size` column.
- **How it helps:** Understanding the typical size of apps can help in resource allocation and development.
- **Business Impact:** Knowing the distribution of app sizes can guide infrastructure planning and optimization of app performance.

4. How many free vs paid apps are there?

- **Hint:** Use `pandas` to count the occurrences of each value in the `Type` column.
- **How it helps:** Gauging the market split between free and paid apps.
- **Business Impact:** Understanding the market dynamics between free and paid apps can inform pricing strategies and marketing campaigns.

5. What is the most common content rating for apps?

- **Hint:** Use `pandas` to find the mode of the `Content Rating` column.
- **How it helps:** Understanding the target audience for most apps.
- **Business Impact:** Knowing the prevalent content rating can help in designing apps that appeal to the largest user base and ensuring compliance with age-appropriate content guidelines.

6. What are the top 5 most installed apps?

- **Hint:** Sort the dataset by the `Installs` column and select the top 5.
- **How it helps:** Identifying popular apps can inform marketing and development strategies.
- **Business Impact:** Highlighting the most installed apps can showcase successful case studies and strategies that can be emulated to achieve similar success.

7. How many apps have a rating of 4.0 and above?

- **Hint:** Use `pandas` to filter the `Rating` column for values >= 4.0.

- **How it helps:** Focusing on high-quality apps can highlight successful development practices.
- **Business Impact:** Identifying high-rated apps can help in understanding best practices and setting benchmarks for new app developments to achieve high user satisfaction.

8. What is the average number of reviews for free vs paid apps?

- **Hint:** Group the data by `Type` and calculate the mean of the `Reviews` column.
- **How it helps:** Assessing user engagement with free versus paid apps.
- **Business Impact:** Understanding the engagement levels can inform promotional strategies and product positioning to maximize user feedback and engagement.

9. What is the average app size for each category?

- **Hint:** Group the data by `Category` and calculate the mean of the `Size` column.
- **How it helps:** Understanding category-specific size requirements can guide development.
- **Business Impact:** Knowing the average size for each category can help optimize resource allocation and improve app performance specific to each category.

10. How many apps were last updated in 2018?

- **Hint:** Convert `Last Updated` to datetime and filter for the year 2018.
- **How it helps:** Analyzing update frequency can inform maintenance strategies.
- **Business Impact:** Understanding update patterns can help in planning regular updates and maintaining app relevance and user engagement over time.

Medium-Level Questions:

1. What is the correlation between the number of installs and the app rating?

- **Hint:** Use `pandas` and `numpy` to calculate the correlation between the `Installs` and `Rating` columns.
- **How it helps:** Understanding the relationship between popularity and user satisfaction.
- **Business Impact:** Identifying if higher installs lead to better ratings can guide marketing and user acquisition strategies.

2. Which app categories have the highest average rating?

- **Hint:** Group the data by `Category` and calculate the mean of the `Rating` column.
- **How it helps:** Identifying top-performing categories based on user ratings.

Note: If you face any difficulty or you want to discuss the query to the mentor feel free to contact the Mentor over LinkedIn - https://www.linkedin.com/in/shivavashishtha/

- **Business Impact:** Focusing on high-rated categories can guide investment and development decisions to enhance user satisfaction.

3. How does the price of an app affect its average rating?

- **Hint:** Group the data by `Price` and calculate the mean of the `Rating` column. Filter for paid apps only.
- **How it helps:** Understanding if more expensive apps tend to have higher or lower ratings.
- **Business Impact:** Informing pricing strategies to optimize user satisfaction and revenue.

4. What is the distribution of app ratings across different content ratings?

- **Hint:** Use `seaborn` to create a boxplot of `Rating` against `Content Rating`.
- **How it helps:** Visualizing how ratings vary with content appropriateness.
- **Business Impact:** Tailoring app content and marketing strategies based on content rating performance.

5. Which genres have the most apps with over 1 million installs?

- **Hint:** Filter the data for apps with `Installs` greater than 1 million and then group by `Genres`.
- **How it helps:** Identifying popular genres among high-install apps.
- **Business Impact:** Guiding genre-specific development and marketing efforts to capitalize on high-install trends.

6. How frequently do apps get updated? Calculate the average time between updates.

- **Hint:** Convert `Last Updated` to datetime and calculate the difference between consecutive updates for each app.
- **How it helps:** Understanding the update cycle of apps.
- **Business Impact:** Planning regular updates to maintain app relevance and user engagement.

7. What is the impact of app size on the number of installs?

- **Hint:** Create a scatter plot using `matplotlib` or `seaborn` with `Size` and `Installs`.
- **How it helps:** Analyzing if app size affects download popularity.
- **Business Impact:** Optimizing app size to balance performance and user appeal.

Note: If you face any difficulty or you want to discuss the query to the mentor feel free to contact the Mentor over LinkedIn - https://www.linkedin.com/in/shivavashishtha/

8. Which apps have the highest number of reviews, and what are their ratings?

- **Hint:** Sort the dataset by `Reviews` and select the top apps, then check their `Rating`.
- **How it helps:** Identifying highly reviewed and rated apps can provide insights into user preferences.
- **Business Impact:** Learning from highly reviewed apps to improve app features and user engagement.

9. How does the content rating distribution differ between free and paid apps?

- **Hint:** Use `pandas` to create a crosstab of `Content Rating` and `Type`.
- **How it helps:** Understanding the target audience for free vs paid apps.
- **Business Impact:** Developing targeted marketing strategies based on content rating and app type.

10. What are the top 5 categories with the most installs?

- **Hint:** Group the data by `Category` and sum the `Installs` column, then sort and select the top 5.
- **How it helps:** Identifying categories with high user demand.
- **Business Impact:** Prioritizing development and marketing efforts in categories with the highest install rates.

Advanced-Level Questions:

1. What are the top 10 apps with the highest ratings, and how do their number of reviews and installs compare?

- **Hint:** Filter the dataset for apps with the highest ratings, then compare their `Reviews` and `Installs`.
- **How it helps:** Understanding if the highest-rated apps also have high user engagement and popularity.
- **Business Impact:** Identifying successful apps to benchmark new app development and marketing strategies.

2. Analyze the trend of app updates over time. Are there any noticeable patterns or seasonal trends?

- **Hint:** Convert `Last Updated` to datetime and plot the frequency of updates over time using `matplotlib` or `seaborn`.

Note: If you face any difficulty or you want to discuss the query to the mentor feel free to contact the Mentor over LinkedIn - https://www.linkedin.com/in/shivavashishtha/

- **How it helps:** Identifying patterns in app updates can reveal industry practices or seasonal effects.
- **Business Impact:** Planning update schedules to align with industry trends and maximize user engagement.

3. How does the average rating of apps change with the number of installs? Create a binned analysis.

- **Hint:** Create bins for `Installs` and calculate the average `Rating` for each bin.
- **How it helps:** Understanding how user ratings vary with app popularity.
- **Business Impact:** Tailoring user acquisition strategies based on how ratings evolve with the user base.

4. Perform sentiment analysis on app reviews (if review text is available) to determine the common themes in high and low-rated apps.

- **Hint:** Use NLP techniques in Python to analyze the sentiment of review texts and categorize themes.
- **How it helps:** Gaining deeper insights into user feedback and common issues or praises.
- **Business Impact:** Addressing common complaints and enhancing features that users love to improve app ratings.

5. What is the relationship between app genre and user ratings? Are certain genres consistently rated higher or lower?

- **Hint:** Group the data by `Genres` and calculate the mean and median of the `Rating` column.
- **How it helps:** Identifying genre-specific trends in user satisfaction.
- **Business Impact:** Focusing on high-rated genres for new app development to align with user preferences.

Additional Considerations for Advanced Questions:

- Ethical and Privacy Concerns: While developing predictive models and handling patient data, it's crucial to consider the ethical implications and ensure privacy and data protection standards are met.
- Interdisciplinary Collaboration: Engage with clinical experts, healthcare providers, and patients to validate findings and refine intervention strategies.
- Continuous Improvement: Consider these analyses as part of an ongoing effort to improve healthcare delivery. Regularly update models and strategies based on new data and outcomes.

Note: If you face any difficulty or you want to discuss the query to the mentor feel free to contact the Mentor over LinkedIn - https://www.linkedin.com/in/shivavashishtha/

Deliverables

- Case Study Document: Includes problem statement, data dictionary, and questions.
- Solution Guide: Detailed answers and explanations for each question.
- Additional Resources: References for further exploration.

Desired Outcome

The trainees will develop an analytical and logical mindset, understanding the importance of various factors in loan analysis. They will learn to apply different data analysis techniques to uncover insights and make data-driven decisions.