* **INTRODUCTION:**

This report summarizes the completion of five tasks related to usage of mobile app analysis, these tasks involved processing and analysing the app data from various categories, generating visualizations and filtering information to meet specific criteria. Main objective of the tasks was to explore app-related insights, in a meaningful way through interactive and static visualizations. By using plotly created dynamic visualizations and dashboards in the python. Where it has helped to explore the various kinds of actions in python by creating different visualizations like

1. Bar chart for top categories
2. Pie Chart Paid apps and free apps percentage
3. Ratings histogram counts
4. Bar chart for sentiment scores minus side & plus side
5. Bar chart installs\_by\_category' to get the top 10 categories by installs
6. Line chart no. of updates as per the year
7. Bar chart for top category generated revenue
8. Bar plot for Top Genres
9. Scatter plot between the last updated and rating
10. Box plot between the type and ratings
11. Dashboard

* **BACKGROUND:**

The dataset contains detailed information about various mobile applications, including app categories, ratings, reviews and other relevant, size, installs, type, price, content rating, generes, Last updated, current version, android version in (PLAY STORE DATA SET). User Reviews Data set contains App Names, Translated Reviews, sentiment, sentiment polarity, sentiment subjectivity. This data set with the following columns helps to achieve the given tasks and in the same way, it’s a greatest learning source, where I was taught some new libraries in python for data manipulation, data cleaning and building some static graphs using plotly visualizations and also taught some main topic known as sentiment analysis for the text related subject. Numerous things can be list down from this internship which helped to me enhance my abilities in complex findings, understanding the of NLP (Natural Language processing), further more this internship helped me sharpen my problem-solving abilities and also preparing me for future challenges related to Data Analytics Field.

* **LEARNING OBJECTIVES:**
* The primary learning objectives of this analysis includes:
* Mastering the use of python and plotly for visualizations.
* Sentiment Analysis for the data which helps to understand the given review is positive, negative, neutral.
* Gaining proficiency of understanding and performing the data analysis with specific filters for targeted insights.
* Learning how to create an interactive visualization in python
* Developing skills in data filtering, cleaning and preprocessing.
* Learning how to create an interactive dashboard in python using plotly visualizations.
* Learned to import various new libraries for plotting the new visualizations in python.
* **ACTIVITIES**

1. Generate a word cloud for the most frequent keywords found in 5-star reviews, but exclude common stop words and app names. Additionally, filter the reviews to include only those from apps in the "Health & Fitness" category.
2. Create a dual-axis chart comparing the average installs and revenue for free vs. paid apps within the top 3 app categories. Apply filters to exclude apps with fewer than 10,000 installs and revenue below $10,000 and android version should be more than 4.0 as well as size should be more than 15M and content rating should be Everyone and app name should not have more than 30 characters including space and special character.
3. Create an interactive choropleth map using Plotly to visualize global installs by country. Apply filters to show data for only the top 5 app categories and highlight countries where the number of installs exceeds 1 million and App category should not start with character “A” , “C” , “G” and “S” . This Graph should not work between 12PMto6PM.
4. Generate a heatmap to show the correlation matrix between installs, ratings, and review counts. Filter the data to include only apps that have been updated within the last year and have at least 100,000 installs and reviews count should be more than 1k and genres name should not be Starting with characters A , F , E , G , I , K.This Graph should work only between 3 pm to 6 pm.
5. Plot a time series line chart to show the trend of total installs over time, segmented by app category. Highlight periods of significant growth by shading the areas under the curve where the increase in installs exceeds 20% month-over-month and content rating should be teen and app name should start with letter ‘E’ and installs should be more than 10k as well as this chart should work between 4 Pm to 8Pm.

* **TASKS: The activities involved working through several complex tasks:**
* Creating a **word cloud from 5-star reviews**, focusing on health-related apps. As the apps\_df and user reviews data sets are both merged first and then later filtering only the 5-star rating apps from health & fitness category, has given zero rows with 17 columns appeared. Ratings present in the apps\_df column has 5-star ratings given for health and fitness category apps. After merging both datasets and applying the filter, there are no 5-star apps in the health & fitness category with reviews in the translated review column. as a result, no 5-star rated apps are available for generating a word cloud from the translated reviews. As I changed the ratings filter to 4.0 and above 2578 rows and 17 columns got reflected, 4.0 to 4.8 translated reviews are present.
* Generating **a dual-axis chart** comparing the installs and revenue of free vs paid apps in top categories. The following conditions were applied to filter the data: Installs: apps should have a size greater than 15 mb
* content rating: the content rating should be "everyone”. Android version: the android version should be greater than 4.0.app name length: the length of the app name should be 30 characters or fewer. Price: apps should have a price greater than $10(for paid apps). outcome of filtering: after applying all the above filters, no data values met the specified requirements. in other words, no apps qualified based on the given conditions.
* **CONCLUSION:** As a result, due to the absence of qualifying data after applying these filters, it is not possible to generate the dual axis chart comparing free vs paid apps. the filtering conditions were too stringent, and no apps in the dataset met the criteria for inclusion.
* Developing an interactive choropleth map to visualize global installs by country, considering multiple filters.
* Executing the third task choropleth for a **choropleth map to work correctly**, it is crucial that the data points are associated with specific geographic locations, typically countries (or regions, states, etc..). the map needs to link the data (such as population) to geographic boundaries. this is usually done using country names (e.g. "united states", "India", "Australia") or country codes (e.g., "us", "in", "br"). These identifiers allow the map software to accurately place the data in the correct geographic area on the map. Without these key pieces of information-country names or country codes-the data cannot be mapped to the corresponding countries, making it impossible to create an accurate choropleth map. essentially, without knowing where the data should be applied geographically, the mapping process cannot proceed.
* **SUMMARY:** The data necessary for building a choropleth map is lacking the crucial geographic identifiers: country names or country codes are essential to link the data with its respective locations on the map. without this linkage, the choropleth map cannot be created because the software will not know where to assign the data geographically.
* **Generating a heatmap** for correlation between app installs, ratings, and review counts. No apps updated in the last year (2023): since the dataset only contains data from 2016 to 2018, the filter to include only apps updated in 2023 doesn't match any records, which means no data is available for this part of the analysis.
* **DATA CLEANING**: The 'reviews' column contained non-numeric values such as commas and plus signs. these had to be cleaned and converted to integers for analysis. we need a recent dataset that includes apps updated within the last year. the heatmap visualization would be effective for examining correlations once the dataset contains apps that meet all the filtering criteria.
* Plotting a **time-series line chart** to track the trend of installs over time, segmented by app category.
* Generated a Plot time series line chart, which works between the 4 pm to 8 pm.
* After cleaning the dataset by removing the duplicates, filtering based on the content rating and app name, and converting the “Installs” column to integers. We ensure data is accurate and suitable for time series analysis
* Month over Month growth metric highlights trends in app installs for each category, allowing us to identify categories with significant growth games and family apps has shown increased 20%.
* Category wise insights different app categories show distinct trends in installs over time with some experiencing consistent growth and others being relatively stable. Reflecting shifts in user interest or market dynamics.
* The time series chart provides a clear visualization of total installs by category, helping to identify long-term trends, seasonal peaks and market shifts in app usage, which can inform business or marketing strategies.
* **SKILLS & COMPETENCIES:**

The tasks involved a range of skills and competencies related to data analysis, visualizations, and python programming:

* **Data Filtering and Preprocessing:** Developing expertise in cleaning and filtering data to meet specific conditions, such as ensuring only apps with certain characteristics were analyzed.
* Visualization Techniques: Mastering the use of python libraries like Plotly for creating
* Interactive visualizations (e.g., heatmaps, choropleth maps, dual-axis charts).
* **Data Analysis:** Analysing app-related data to identify trends and insights, especially focusing on metadata like app ratings, installs and reviews.
* **Sentiment Analysis:** Using sentiment analysis to determine whether reviews are positive, negative or neutral enhancing understanding of user feedback.
* Interactive Dashboards: Learning to create interactive dashboards that allow users to explore the data dynamically.
* **FEEDBACK AND EVIDENCE:** Feedback was mainly based on the execution of specific tasks and their outcomes:
* **EVIDENCE:** Working file created in the spyder for each task.
* **WORD CLOUD:** After filtering for 5-star health & fitness apps, no data was available to generate a word cloud. Adjusting the filter to include apps with ratings of 4.0 and above led to 2,578 rows of translated reviews.
* **DUAL-AXIS CHART:** Due to overly strict filtering conditions, no data met the criteria for comparing installs and revenue for free vs paid apps, highlighting the importance of balanced filtering.
* **CHOROPLETH MAP:** The feedback from attempting to generate a choropleth map Indicated the need for geographic identifiers (e.g., country names or country codes), which were lacking in the dataset, preventing the successful creation of the map.
* **HEATMAP:** No apps were updated in the last year (2023), and the reviews column required cleaning before a heatmap could be generated.
* **TIME SERIES LINE CHART:** The time series line chart demonstrates the month-over-month growth of over 20% in the Family and Game categories, with the analysis restricted to the period between 4 PM and 8 PM. The data was filtered by category, duplicates were removed, and the month and year were extracted to categorize the installs over time.
* **CHALLENGES AND SOLUTIONS:**
* Several Challenges were encountered during the tasks, each with its respective solutions:
* **CHALLENGE: LACK OF APPS UPDATED IN THE LAST YEAR (2023)**
* **Solution:** The dataset only contained apps from 2016 to 2018, so no apps met the condition for being updated in 2023. The solution would be to use a more recent dataset for analysis.
* **CHALLENGE: STRICT FILTERING CONDITIONS FOR THE DUAL-AXIS CHART**
* **Solution:** The filtering conditions for installs, content rating, android version, and app size were too restrictive, resulting is no qualifying apps.
* **CHALLENGE: MISSING GEOGRAPHIC IDENTIFIERS FOR CHOROPLETH MAP**
* **Solution:** The dataset lacked crucial geographic information, such as country names or codes. Adding these identifiers to the dataset would resolve the issue and allow the choropleth map to be created.
* **CHALLENGE: DATA CLEANING FOR THE HEATMAP**
* **Solution:** The reviews column had non-numeric values (commas and plus signs), which had to be cleaned and converted to integers. After cleaning the data, the heatmap could have been generated once the appropriate apps were included.
* **CHALLENGE:** **FOR THE 5th TASK**
* **Challenge1:** Filtering apps by specific criteria, such as starting with the letter "E" and having installs above 10,000, can narrow the dataset significantly, making it harder to generate insights for other apps.  
  **Solution:** Applying precise filtering conditions ensures the dataset is focused on relevant apps, allowing for more targeted analysis while preventing unnecessary data overload.
* **Challenge 2:** Sorting categories and apps for the time series chart can become complicated, particularly when dealing with a large number of entries.  
  **Solution:** Proper sorting by category and update date streamlines the data, ensuring an organized timeline and making the chart easier to interpret.
* **Challenge 3:** Restricting the chart's execution to a specific time window (4 PM to 8 PM) adds complexity to the code and may limit automated execution.  
  **Solution:** Implementing time-based logic (datetime.now().hour) ensures that the script runs only during the desired time window, preventing unnecessary execution outside of this period.
* **OUTCOMES & IMPACTS:**
* Improved Data Cleaning Skills: The process of cleaning data, such as converting reviews to integers, helped strengthen the ability to handle real-world datasets with inconsistencies.
* Data Analysis Insights: Although some tasks didn’t produce the expected results due to filtering issues, the tasks helped developed a deeper understanding of how to filter data derive meaningful insights.
* Visualization Experience: The experience with plotly and interactive visualizations has improved the ability to create engaging informative visualizations for large datasets.
* The time series line chart is configured to operate with in the time frame of 4 pm to 8 pm. This involved importing the current date and ensuring that the graph only functions within the specified time window.
* **CONCLUSION:** The completion of these tasks provided valuable lessons in data filtering, analysis, and visualization, while also highlighting the importance of data quality and accuracy. Some tasks, such as the creation of the dual-axis chart and choropleth map, word cloud, heat map faced limitations due to overly restrictive filters or missing data. However, the time series plot chart has been come out well with the increase in 20% over month by month in categories of family and gaming related apps. The overall process demonstrated the need for continuous data cleaning, preprocessing, and flexibility in task design. The insights gained from these tasks will be applied to future analyses, especially by utilizing more up-to-date datasets and adjusting filtering criteria to ensure meaningful results.