INTERNSHIP PROJECT REPORT

Project Title: Real-Time Google Play Store Data Analysis with Python

Submitted by:

Mishita Maggo

NULLCLASS

Issued May 11 2025

CERTIFICATE OF TRAINING

THIS CERTIFICATE IS PRESENTED TO

Mishita Maggo

has successfully completed online training on

Data Analyst

and real-time project training on

Learn to Build Real time Google Play store data analytics - python

Vetriselvan G.
CEO

STIN: 33AAICN0803F1ZZ
N: U80301TZ2022PTC038231
OC ID: 6820ca79b69ed9cb7aec1f76

INDEX

INTRODUCTION

In the current digital age, mobile applications are essential for influencing business goals and customer behaviour. One of the biggest marketplaces for Android apps, the Google Play Store offers millions of apps in a variety of categories. It has become crucial for developers, marketers, and analysts to use data analysis to understand user preferences, app performance, and market trends.

In this internship project, real-time analysis of Google Play Store data is done with Python and its robust data manipulation modules, including NumPy and Pandas. Getting useful insights from publicly accessible app metadata, such as user ratings, download counts, app categories, and more, is the goal. The goal of this study is to find correlations, patterns, and trends that might help improve app development and marketing decision-making.

Because of its extensive ecosystem of data science tools and versatility, Python forms the foundation of this project. While Pandas offers robust data structures for managing and analysing big datasets, NumPy is used for effective numerical computations. These tools work together to effectively clean, convert, and visualise data, which makes it easier to gain a thorough grasp of the Play Store ecosystem.

This study describes the project's approach, tools, analytical methods, and main conclusions. It shows how real-time data analysis may be utilised to obtain useful insights into the ever-changing field of mobile applications.

BACKGROUND

With millions of apps available on stores like the Google Play Store, the mobile app market is expanding at a never-before-seen pace. Actionable insights are necessary for organisations and developers to enhance their apps, spot user patterns, and maintain their competitiveness. When properly examined, raw data from the Play Store, including ratings, reviews, download counts, and category classification, can provide insightful information.

With the help of NullClass, this internship gave me the chance to work with real-time data from the Google Play Store and learn how to use Python and its libraries, such NumPy and Pandas, to clean, analyse, and visualise it. The objective was to investigate how data science approaches may improve decision-making for marketers and app developers and assist in deriving significant conclusions from massive datasets.

LEARNING OBJECTIVES

The internship was designed to foster the growth of both technical and analytical abilities. One of the main learning goals was to:

Gaining knowledge of data science workflows entails knowing how to collect, process, and analyse vast amounts of real-time data.

Python hands-on programming: creating clear, effective code to work with datasets.

Knowledge of NumPy and Pandas: Utilising NumPy for effective array processing and Pandas for dataframes.

Finding patterns, anomalies, and connections in metrics pertaining to apps is known as exploratory data analysis, or EDA.

Visualisation Skills: Making use of dashboards, charts, and graphs to display the results.

Report Writing: Summarizing the entire process and communicating insights.

ACTIVITIES AND TASKS

Various technical and analytical tasks were completed during the internship, such as:

Data acquisition: Used validated data sources and scraping tools (like BeautifulSoup and requests) to source real-time data. characteristics such as program names, developer details, rating, quantity of reviews, size, installs, kind (free/paid), content rating, and genre were the main focus.

Data Preparation: eliminated redundant records, dealt with erroneous or missing data (such as apps without install counts or ratings).

Strings were converted to numerical representations (for example, installing counts such as "1,000,000+" to integers).

Cleaning and Normative Data: standardised app sizes and categories. used custom lambda functions, fillna(), astype(), and groupby(), among other Pandas actions.

EDA, or exploratory data analysis: examined the most downloaded and highly rated apps, compared paid and free apps in many categories, and determined patterns by comparing the size of the app to its rating or installations.

Visualisation: used libraries like seaborn and matplotlib to create plots, displayed correlation heatmaps, rating histograms, and category distribution.

SKILLS & COMPETENCIES

Technical Proficiency: Python Programming: Composing organised, effective Python data processing programs.

NumPy: Effective management of numerical data and arrays.

Pandas: Proficiency with DataFrames, filtering, grouping, and data processing.

Making educational visual representations of data is known as data visualisation.

Web scraping: Being aware of the fundamental methods for gathering real-time data.

Analytical Skills: Deciphering big data sets to identify patterns and irregularities.

utilising real-world data to derive relevant insights.

creating and evaluating theories based on data from statistics.

Soft Skills: Time Management: Fulfilling assignments and deliverables by the due date.

Solving problems: debugging programming and fixing discrepancies in data.

Communication: Presenting results and recording procedures.

FEEDBACK & EVIDENCE

During the internship, mentors and team members at NullClass provided comments. The quality of the work and the clarity of the ideas were enhanced by the helpful feedback that was given during weekly checkpoints and reviews.

Notebook:
https://colab.research.google.com/drive/127ETLFaone80yF1Sdfuwgzyc3ReuLltX?usp=drive_link

Task 1: Scatter Plot – Revenue vs Installs (Paid Apps Only)

```
[12] # Clean the data
     data['Installs'] = data['Installs']
     # Check if 'Price' column is of type object (string) before applying str.replace
     if data['Price'].dtype == object:
         data['Price'] = data['Price'].str.replace('$', '', regex=True).astype(float)
     else:
         # If 'Price' is already numeric, handle potential '$' signs differently
         # (e.g., if '$' is present, treat as string and apply str.replace, otherwise, leave as is)
         data['Price'] = pd.to_numeric(data['Price'].astype(str).str.replace('$', '', regex=True), errors='coerce')

     data['Revenue'] = data['Price'] * data['Installs']

     # Filter paid apps
     paid_apps = data[(data['Price'] == 'Paid').notna() & data['Installs'].notna() & data['Category'].notna()]

     # Create the scatter plot
     fig = px.scatter(
         paid_apps,
         x='Installs',
         y='Price',
         color='Category',
         trendline='ols',
         title='Price vs Installs for Paid Apps'
```

CHALLENGES AND SOLUTIONS

Challenges:
Data Inconsistency: Missing values or irregular formats were common in real-time data.

Parsing problems: Having trouble cleaning up complicated fields, such as install counts that contain symbols or units.

Performance bottlenecks: There have occasionally been performance lags while working with huge datasets.

Limitations on Scraping: Restrictions on scraping because of APIs or website security measures.

Solutions: Applied sophisticated Pandas cleaning methods (such as chaining processes).

Fields were cleaned and standardised using regular expressions and string manipulation.

NumPy arrays and Pandas'.loc[] procedures were used to optimise data management.

When scraping was not possible, use static datasets or, if available, APIs.

OUTCOMES AND IMPACT

There were notable practical and educational results from the project:

gained a basic understanding of the data analysis process, from collection to interpretation.

gained expertise handling big, disorganised real-world data.

developed a compact yet effective analytical framework that can be applied to data from comparable apps.

developed visual summaries and dashboards that showcase important Play Store app metrics.

positioned to use similar abilities in other fields, such social networking, e-commerce, or health data.

My internship gave me practical experience with the types of problems that data analysts encounter and gave me the skills I needed to solve them on my own.

CONCLUSION

An invaluable educational experience that helped close the gap between theory and practice was the internship at NullClass. Using Python, NumPy, and Pandas, I gained essential skills in data collecting, cleaning, analysis, and visualisation while working with real-time Google Play Store data. I gained more technical proficiency and a deeper comprehension of the mobile application ecosystem thanks to the insights this investigation produced.

I am now more comfortable tackling challenging data-driven problems using methodical tools and structured thinking, and this project established a strong basis for my future work in data science and analytics.