Assignment-1

By-Rewant Sharma C0894265

Note:code attached

**Flight Prices:**

## Q1. Load the flight price dataset and examine its dimensions. How many rows and columns does the dataset have?

A screen shot of a computer

Description automatically generated

This prints the head and the tail of the data .

This prints the rows and columns of the data.

A screenshot of a computer

Description automatically generated

## Q2. What is the distribution of flight prices in the dataset? Create a histogram to visualize the distribution.

A screen shot of a computer

Description automatically generated

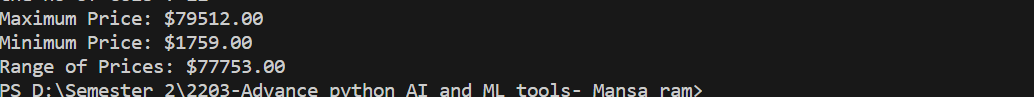
A screenshot of a computer screen

Description automatically generated

## Q3. What is the range of prices in the dataset? What is the minimum and maximum price?

A black screen with colorful text

Description automatically generated



## Q4. How does the price of flights vary by airline? Create a boxplot to compare the prices of different airlines.

A screen shot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

## Q5. Are there any outliers in the dataset? Identify any potential outliers using a boxplot and describe how they may impact your analysis.

A group of small squares

Description automatically generated with medium confidence



Indeed, there are some outliers in the data set. For instance, all of the data in the arrival time column is on time, but some of the data have timestamps like 01:10 on March 22, as illustrated below.

A screenshot of a computer program

Description automatically generated

You can find any spots that are noticeably above or below the whiskers by looking at the boxplot. These are possible outliers that could distort the distribution and have an effect on your research.   
The statistical measure, data visualisation, model performance, and data interpretation can all be impacted by outliers.

A screenshot of a computer

Description automatically generated

## Q6. You are working for a travel agency, and your boss has asked you to analyze the Flight Price dataset to identify the peak travel season. What features would you analyze to identify the peak season, and how would you present your findings to your boss?

I looked at the following traits in order to determine the season of travel in the airline ticket dataset:  
Time or date: Finding periods when prices are often higher or lower can be accomplished by examining airline tickets from various months or seasons. This could entail computing the average price for each period by aggregating the data by month or season.  
  
Destination: You can determine the peak travel season by learning which places are more popular at particular periods of the year. Holidays, festivals, or particular weather conditions may increase the popularity of some locations.

Airline: Seasonal variations and pricing strategies may vary throughout airlines. Peak travel times can be determined by examining the average fares provided by several airlines over a period of time.   
Special Events: Peak travel seasons can also be determined by identifying special events, such as festivals, conferences, or significant holidays, that fall during periods of higher travel demand.

## Q7. You are a data analyst for a flight booking website, and you have been asked to analyze the Flight Price dataset to identify any trends in flight prices. What features would you analyze to identify these trends, and what visualizations would you use to present your findings to your team?

1. **Airline Analysis**:
   * Explore price variations across different airlines.
   * Use bar charts or box plots to show average ticket prices per airline.
2. **Source and Destination Cities**:
   * Investigate how prices vary based on departure and arrival cities.
   * Visualize using heatmaps or scatter plots.
   * Consider a geographical map with color-coded routes and prices.
3. **Departure and Arrival Times**:
   * Group flights by time of day (morning, afternoon, evening).
   * Examine price distributions.
   * Use line charts or violin plots.
4. **Stops**:
   * Compare average prices for direct flights vs. layovers.
   * Display results in a bar chart.
5. **Class (Business vs. Economy)**:
   * Analyze ticket prices for business and economy classes.
   * Create side-by-side box plots or grouped bar charts.
6. **Duration vs. Price**:
   * Understand the relationship between flight duration and price.
   * Use scatter plots or regression lines.
7. **Days Left**:
   * Study price changes as departure date approaches.
   * Visualize with a line chart.
8. **Price Trends Over Time**:
   * Create a time series plot to show overall price trends (monthly or seasonal).
9. **Feature Correlations**:
   * Calculate correlations
   * Display a correlation matrix heatmap.
10. **Identify Outliers**:
    * Detect extreme price outliers.
    * Use box plots or scatter plots

I use visuals like this to show the group my findings:

1. Heatmaps,
2. Boxplots,
3. Bar Charts,
4. Time Series Charts

## Q8:You are a data scientist working for an airline company, and you have been asked to analyze the Flight Price dataset to identify the factors that affect flight prices. What features would you analyze to identify these factors, and how would you present your findings to the management team?

As a data scientist analyzing the **Flight Price dataset**, I would focus on several critical features to identify factors affecting flight prices. Here’s a concise breakdown of the relevant factors and how I would present my findings to the management team:

1. **Airline Analysis**:
   * Investigate how prices vary across different airlines.
   * Use bar charts or box plots to show average ticket prices per airline.
   * Present findings: A visual comparison of airline prices, highlighting any significant disparities.
2. **Booking Lead Time**:
   * Analyze how prices change based on the number of days before departure.
   * Create a line chart showing price fluctuations as booking lead time decreases.
   * Present findings: Show the impact of last-minute bookings on ticket prices.
3. **Departure and Arrival Times**:
   * Explore how prices vary with departure and arrival times (e.g., morning, afternoon, evening).
   * Visualize using line charts or scatter plots.
   * Present findings: Highlight peak travel times and their influence on prices.
4. **Source and Destination Cities**:
   * Investigate price variations based on departure and arrival cities.
   * Use heatmaps or scatter plots to visualize these relationships.
   * Present findings: Identify popular routes and their associated price ranges.
5. **Flight Duration**:
   * Understand how flight duration affects prices.
   * Scatter plots or regression analysis can reveal patterns.
   * Present findings: Discuss the trade-off between flight duration and cost.
6. **Number of Stops**:
   * Compare prices for direct flights vs. those with layovers.
   * Use bar charts or box plots.
   * Present findings: Highlight cost differences based on stopovers.
7. **Class (Economy vs. Business)**:
   * Analyze ticket prices for economy and business classes.
   * Side-by-side box plots or grouped bar charts can illustrate this.
   * Present findings: Discuss the premium associated with business class.
8. **External Factors**:
   * Consider external influences such as political events, natural disasters, and economic conditions.
   * Present findings: Explain how external events affect ticket prices.
   * How these factors affect demand and supply.
9. **Correlations**:
   * Calculate correlations between features (e.g., distance, flight time, stops).
   * Use a correlation matrix heatmap.
   * Present findings: Identify strong relationships affecting prices.
10. **Outliers**:
    * Detect extreme price outliers.
    * Use box plots or scatter plots.
    * Present findings: Address any anomalies that significantly impact pricing.

**Presentation Approach**:

* **Executive Summary**: Begin with a concise overview of findings.
* **Visualizations**: Use interactive dashboards (e.g., Tableau) to showcase trends visually.
* **Narrative**: Explain insights in plain language, emphasizing actionable takeaways.
* **Recommendations**: Provide actionable recommendations based on the analysis.
* **Q&A Session**: Engage with the management team to address specific queries.
* **Powerpoint presentations** :these presentations would show trends and many variables affecting the prices

## 

CODE attached above

# Google play store Data

**Q9. Load the Google Playstore dataset and examine its dimensions. How many rows and columns does the dataset have?**

Data cleaning for google data after loading it

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Rows and columns for the datasetA black and white text on a black background

Description automatically generated

## Q10. How does the rating of apps vary by category? Create a boxplot to compare the ratings of different app categories.

A screenshot of a computer

Description automatically generated

## Q11. Are there any missing values in the dataset? Identify any missing values and describe how they may impact your analysis

Yes as provided in the below figure we can see so many null values exist under ratings column.

Using the below code we have identified all the missing values in the data provided

A black rectangular object with a black border

Description automatically generated

Given below are the problems which one would face while working on a missing values dataset

1. **Loss of Information**:
   * Missing values can lead to a loss of valuable information. If a significant portion of data is missing, it may affect the overall quality of your analysis.
   * For example, if you’re analyzing customer data and important features like age or income have missing values, your insights may be incomplete.
2. **Biased Analysis**:
   * When missing values are not handled properly, they can introduce bias into your analysis.
   * For instance, if you ignore missing values in a survey dataset, your conclusions may not accurately represent the entire population.
3. **Incorrect Statistical Measures**:
   * Descriptive statistics (mean, median, etc.) calculated without handling missing values can be misleading.
   * For instance, calculating the average income without addressing missing income values may not reflect the true central tendency.
4. **Model Performance**:
   * Machine learning models can be sensitive to missing data. Some algorithms may not handle missing values well.
   * Imputing incorrect values or ignoring missing data can impact model accuracy and generalization.

## Q12. What is the relationship between the size of an app and its rating? Create a scatter plot to visualize the relationship.

A black screen with text

Description automatically generated

A blue dotted pattern on a white background

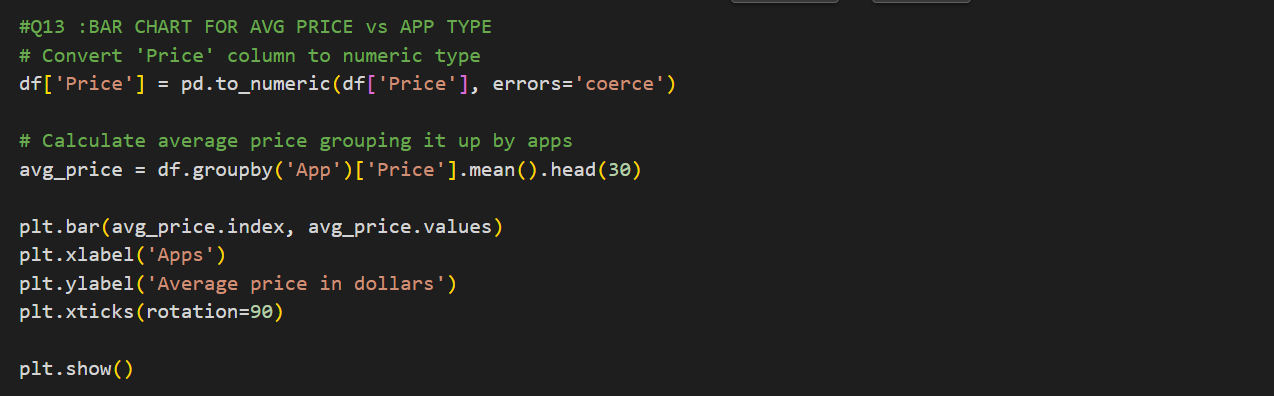
Description automatically generated



A screen shot of a computer

Description automatically generated

## Q13. How does the type of app affect its price? Create a bar chart to compare average prices by app type.



A graph with blue bars

Description automatically generated

A graph with blue and white text

Description automatically generated

## Q14. What are the top 10 most popular apps in the dataset? Create a frequency table to identify the apps with the highest number of installs.

A screenshot of a computer program

Description automatically generated

## Q15. A company wants to launch a new app on the Google Playstore and has asked you to analyze the Google Playstore dataset to identify the most popular app categories. How would you approach this task, and what features would you analyze to make recommendations to the company?

1.**Data Exploration**:

Load the dataset and examine its structure first. Recognise the columns and the kinds of available data. Look for any missing values and deal with them accordingly, by imputing or deleting them.

2.**Data Cleaning:**

Correct any errors in the data, such as misspellings, duplicates, or outliers.   
Make sure the pertinent columns are the right kind of data so they may be analysed.

3.**Feature Selection**:

Determine which elements are important in gauging an app category's popularity. These could consist of:   
Category of apps (e.g., Social, Games, Education)

4. **Data Aggregation:**

To determine metrics like average installs, average ratings, and total reviews for each app category, aggregate the data by app category.

5.**Analysis:**

Determine which app categories are the most popular by analysing the aggregated data according to different parameters. As an illustration:   
Ascertain which categories have the highest average quantity of ratings or installs.   
Look for app categories that have a lot of apps but relatively low average installs or ratings; these could point to a less crowded market.   
Examine how popularity has changed over time by comparing various eras to gain an understanding of patterns.

6.**Visualization:**

Use suitable graphs and charts, such as pie charts, bar charts, and scatter plots, to visually represent the results. This will make it easier for stakeholders to comprehend the data-driven insights.   
Suggestions:

## Q16. A mobile app development company wants to analyze the Google Playstore dataset to identify the most successful app developers. What features would you analyze to make recommendations to the company, and what data visualizations would you use to present your findings?

To identify the most successful app developers in the Google Play Store dataset, I would analyze the following features:

**Number of Downloads/Installs**:

Applications that have received a lot of downloads or installations are successful and well-liked.

**Ratings and Reviews:**

Positive reviews and higher ratings indicate that users are happy and actively using the app.

**App Updates:**

Frequent updates show the developers' dedication to enhancing the software and taking user comments into consideration.

**App Size:**

Smaller app sizes are generally preferred by users, potentially leading to higher adoption rates.

**Price:**

Applications that are free or inexpensive might draw in more users, whereas paid apps might suggest better or more specialised content.

**Retention Rate:**

Apps with high rates of user retention have a good chance of long-term success.   
The following is what I would use for data visualisations:

Visualisations as given below which will be used to show the findings.

**Bar Charts:**

To compare the number of downloads/installs, ratings, and reviews for different developers.

**Scatter Plots:**

To visualize the relationship between app size, ratings, and number of installs.

**Pie Charts:**

To show the distribution of free vs. paid apps by each developer.

**Line Charts:**

To track the frequency of app updates over time for each developer

**Heat maps and word clouds** can also be used.

## Q17. A marketing research firm wants to analyze the Google Playstore dataset to identify the best time to launch a new app. What features would you analyze to make recommendations to the company, and what data visualizations would you use to present your findings?

### Features to Analyze:

1. **Installs by Time**:
   * Examine the ways in which app installations change over time (months, days of the week, even hours).   
      Examine trends to determine the times of peak installation.
2. **Ratings and Reviews Over Time**:
   * Analyse the changes in app reviews and ratings over various time periods.
   * Keep an eye out for trends pertaining to marketing initiatives, app updates, or seasonal occasions.
3. **App Updates Frequency**:
   * Find out how frequently updates are released for popular apps.
   * Applications with regular updates typically function better.
4. **Category-Specific Trends**:
   * Check to see if any app categories have launch windows that are unique.
   * For instance, because people tend to make fitness-related resolutions at the start of the year, fitness applications may be popular.
5. **Seasonal Trends**:
   * Take into account festivals, holidays, and unique occasions.
   * Showcase app installations for particular occasions, such holidays or seasons.

### Data Visualizations:

1. **Line Charts**:
   * Plot the number of app installations, ratings, and reviews over time.
   * Highlight significant events or changes.
2. **Heatmaps**:
   * Show app installations by day of the week and time of day.
   * Identify peak hours for app downloads.
3. **Bar Charts**:
   * Compare average ratings or review counts across different months or seasons.
   * Highlight any spikes or dips.
4. **Box Plots**:
   * Visualize the distribution of app installations or ratings by month.
   * Identify outliers or variations.
5. **Calendar Heatmaps**:
   * Display app installations on a calendar grid.
   * Easily spot patterns related to specific days or months.
6. **Stacked Area Charts**:
   * Show cumulative installations over time, segmented by app category.