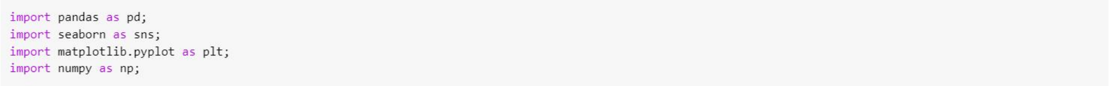
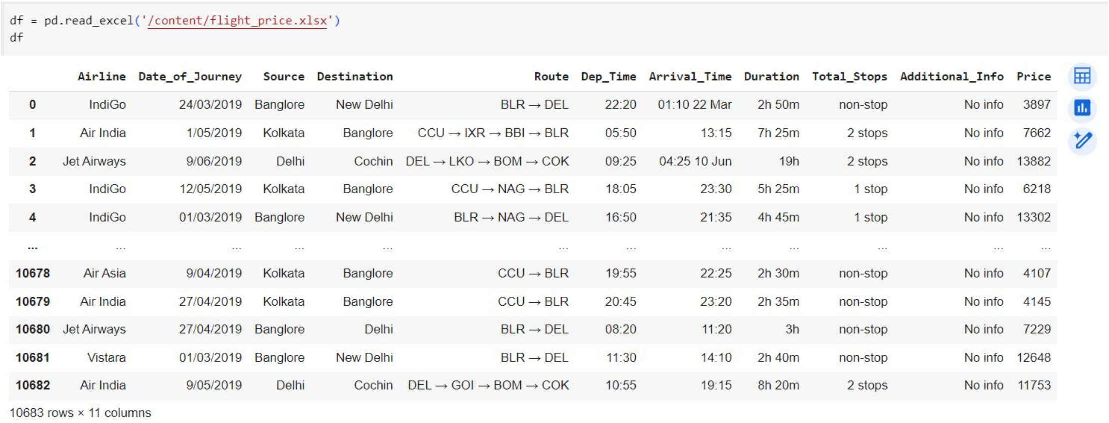
Question 1



```
import pandas as pd;
import seaborn as sns;
import matplotlib.pyplot as plt;
import numpy as np;
```



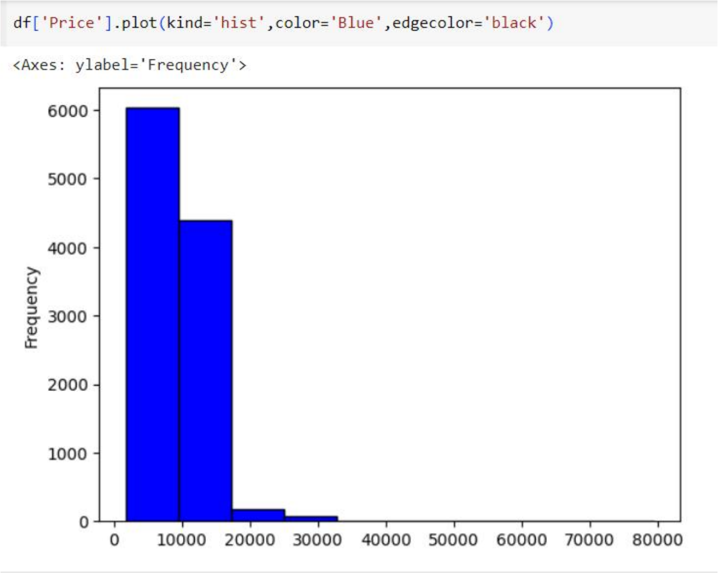
```
df = pd.read_excel('/content/flight_price.xlsx')
df
```

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
...	...	...	...	...	...	...	...	...	...	...	...
10678	Air Asia	9/04/2019	Kolkata	Banglore	CCU → BLR	19:55	22:25	2h 30m	non-stop	No info	4107
10679	Air India	27/04/2019	Kolkata	Banglore	CCU → BLR	20:45	23:20	2h 35m	non-stop	No info	4145
10680	Jet Airways	27/04/2019	Banglore	Delhi	BLR → DEL	08:20	11:20	3h	non-stop	No info	7229
10681	Vistara	01/03/2019	Banglore	New Delhi	BLR → DEL	11:30	14:10	2h 40m	non-stop	No info	12648
10682	Air India	9/05/2019	Delhi	Cochin	DEL → GOI → BOM → COK	10:55	19:15	8h 20m	2 stops	No info	11753

10683 rows × 11 columns

The sheet has 10683 rows and 11 columns

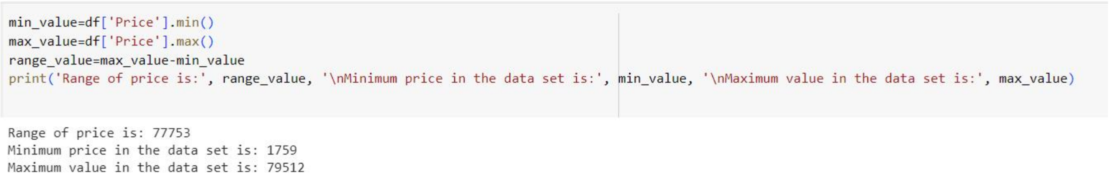
Question 2



```
df['Price'].plot(kind='hist',color='Blue',edgecolor='black')
```

```
<Axes: ylabel='Frequency'>
```

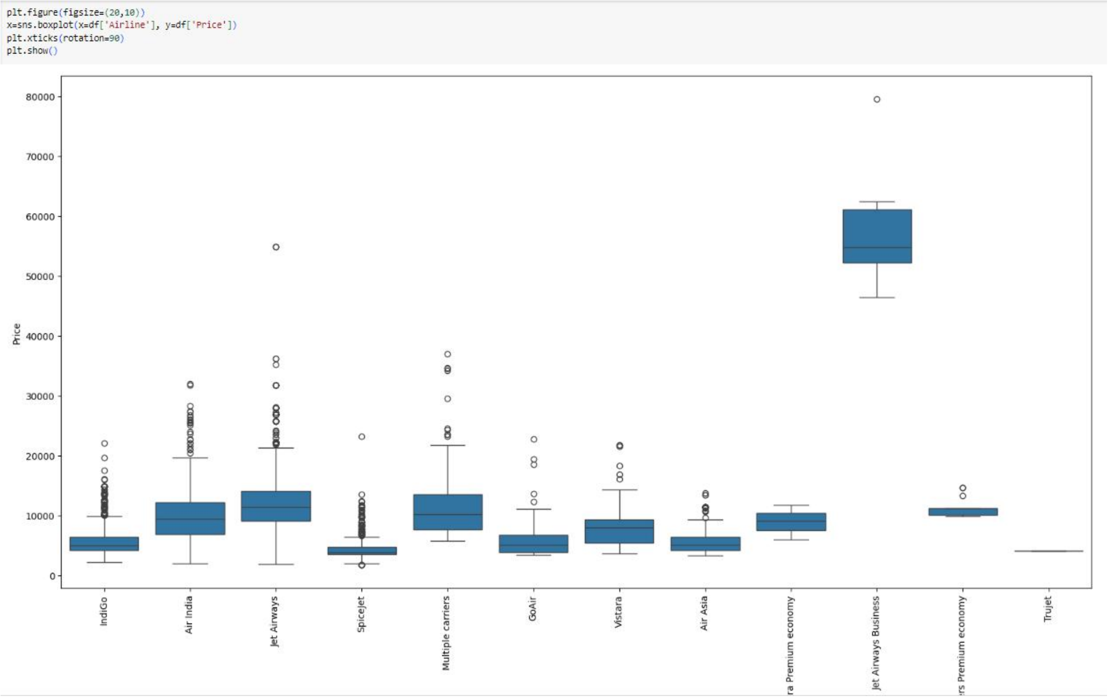
Question 3



```
min_value=df['Price'].min()
max_value=df['Price'].max()
range_value=max_value-min_value
print('Range of price is:', range_value, '\nMinimum price in the data set is:', min_value, '\nMaximum value in the data set is:', max_value)
```

```
Range of price is: 77753
Minimum price in the data set is: 1759
Maximum value in the data set is: 79512
```

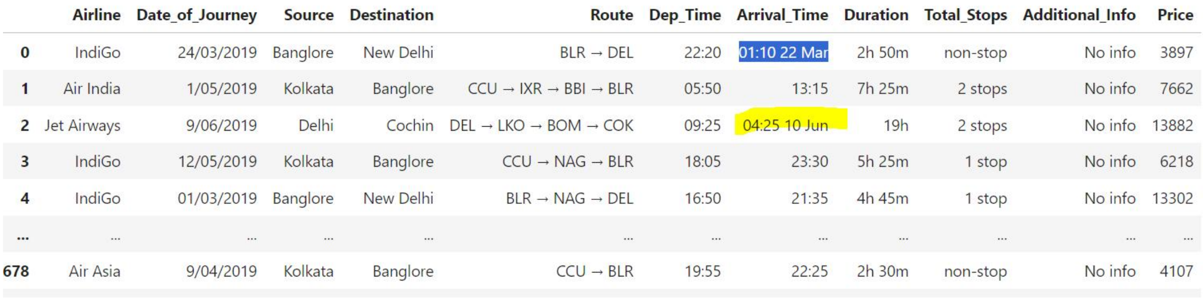
Question 4



```
plt.figure(figsize=(20,10))
x=sns.boxplot(x=df['Airline'], y=df['Price'])
plt.xticks(rotation=90)
plt.show()
```

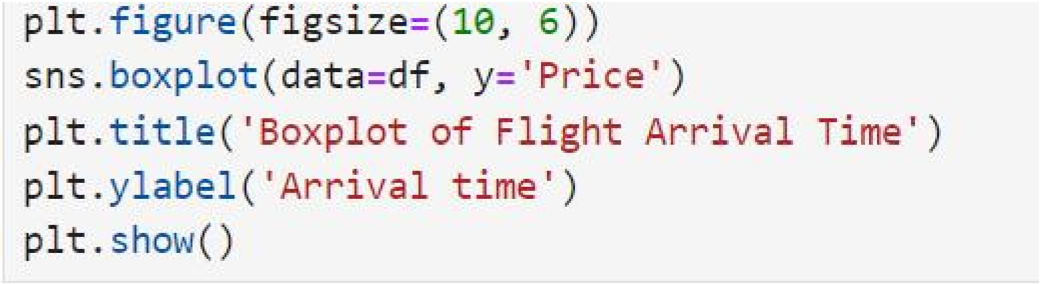
Question 5

Ans:- Yes there is some outliers in the data set that as for example in the arrival time every data is in time but some of the data have time 01:10 22 Mar as shown below

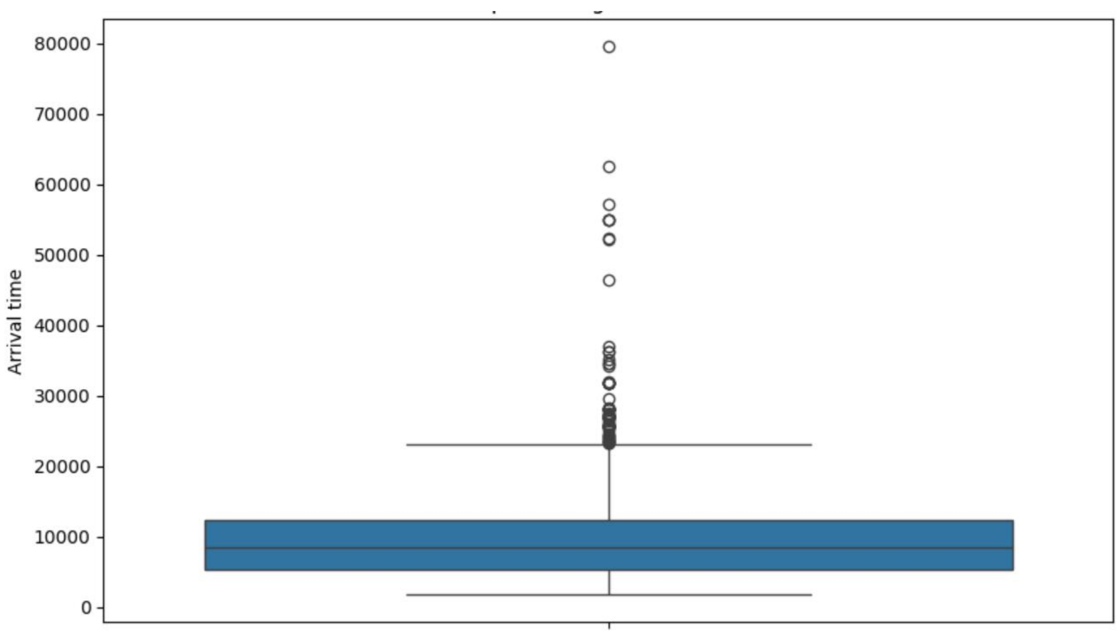


	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
...	...	...	...	...	...	...	...	...	...	...	...
678	Air Asia	9/04/2019	Kolkata	Banglore	CCU → BLR	19:55	22:25	2h 30m	non-stop	No info	4107

It can be shown with the help of boxplot



```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, y='Price')
plt.title('Boxplot of Flight Arrival Time')
plt.ylabel('Arrival time')
plt.show()
```



By examining the boxplot, you can identify any points that lie significantly above or below the whiskers. These points are potential outliers that may skew the distribution and impact your analysis. Outliers can impact the Statistical Measure, Data visualization, Model performance and Data Interpretation.

Question 6
Ans:-To identify trends in flight prices using the Flight Price dataset, I analyzed key features such as:

1.Date or Time: Examining flight prices over different months or seasons to identify price fluctuations throughout the year.
2.Route: Analyzing prices for different routes to understand variations in pricing based on destination and demand.
3.Airline: Investigating average fares offered by different airlines to identify pricing strategies and trends.
4.Booking Class: Understanding pricing differences between economy, business, and first-class bookings.

Question 7
1.Time Series Charts: Showing the trend of flight prices over time, potentially grouped by month or season, to visualize seasonal fluctuations.
2.Bar Charts: Comparing average flight prices across different routes, airlines, or booking classes to identify significant differences.
3.Boxplots: Displaying the distribution of flight prices for different routes or airlines, highlighting outliers and variability.
4.Heatmaps: Visualizing average flight prices across routes and time periods to identify peak travel seasons and popular destinations.

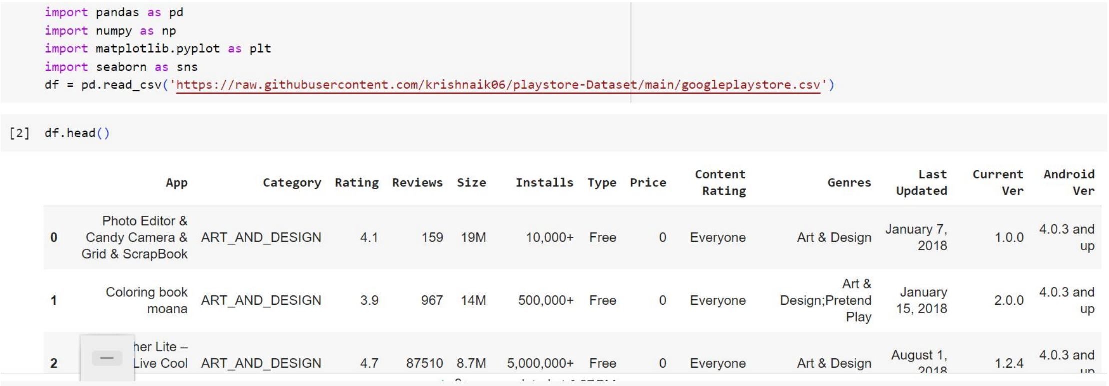
By analyzing these features and utilizing appropriate visualizations, we can effectively identify trends in flight prices and present actionable insights to the team for decision-making.

Question 8
Ans:- o identify factors affecting flight prices in the Flight Price dataset, I would analyze features such as:
Date and Tim, Route, Airline, Booking Class, Advance Booking Period and many more.

I would present my findings to the management team through a combination of:
Visualizations: Utilize graphs, charts, and heatmaps to visually represent relationships between flight prices and the identified factors.
Summary Statistics: Provide key statistical metrics such as mean, median, and standard deviation to summarize the data.
Insights and Recommendations: Offer actionable insights and strategic recommendations based on the analysis to optimize pricing strategies and enhance competitiveness.

Question 1-8.ipynb

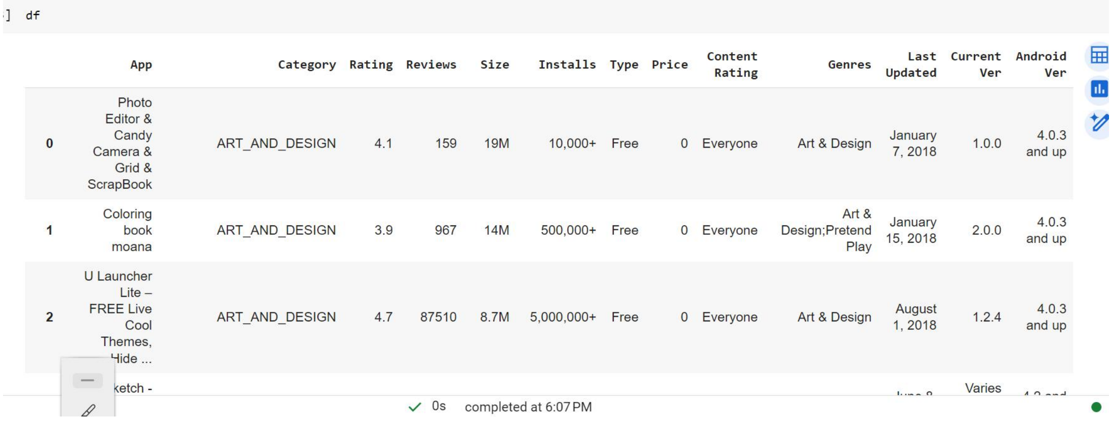
Question 9



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('https://raw.githubusercontent.com/krishnaik06/playstore-Dataset/main/googleplaystore.csv')
```

```
[2] df.head()
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	her Lite – Live Cool	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up

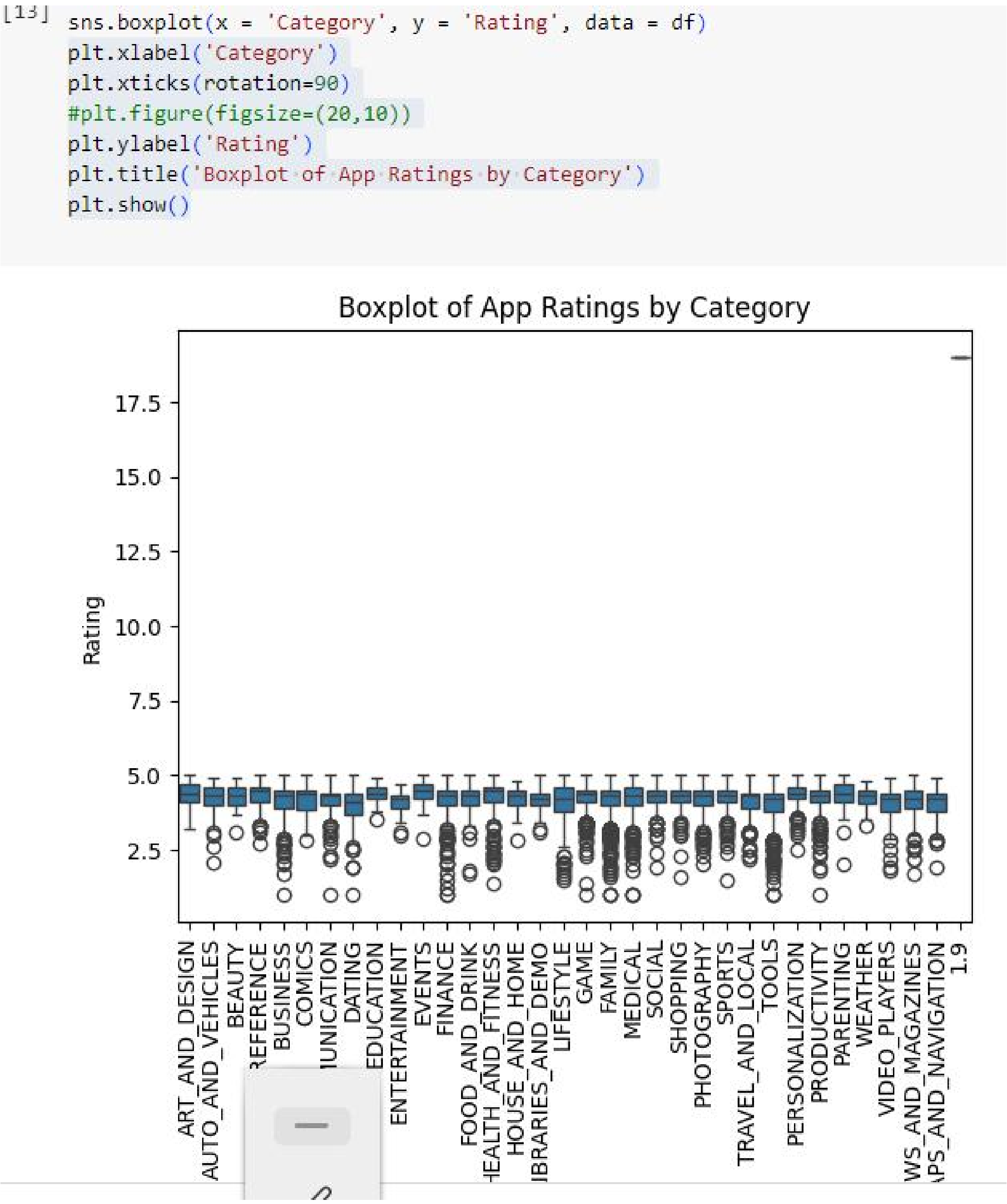


```
] df
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
	ketch -											Varies	

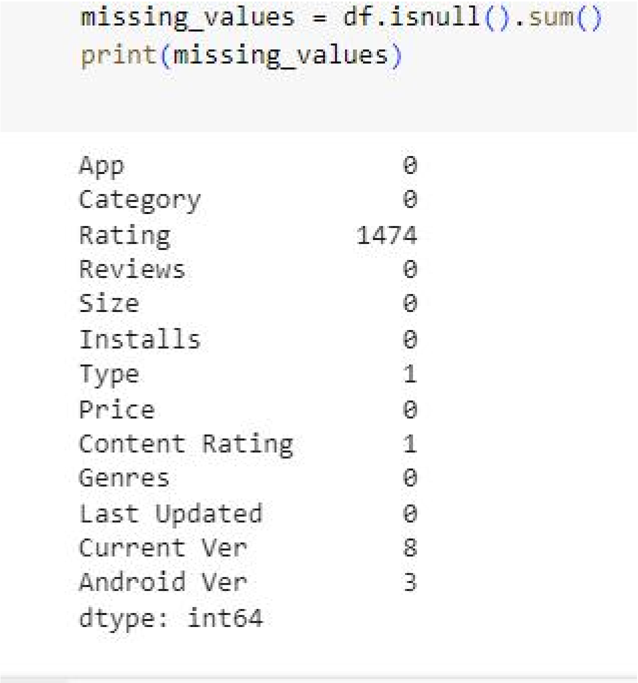
✓ 0s completed at 6:07 PM

Question 10



```
sns.boxplot(x = 'Category', y = 'Rating', data = df)
plt.xlabel('Category')
plt.xticks(rotation=90)
#plt.figure(figsize=(20,10))
plt.ylabel('Rating')
plt.title('Boxplot of App Ratings by Category')
plt.show()
```

Question 11



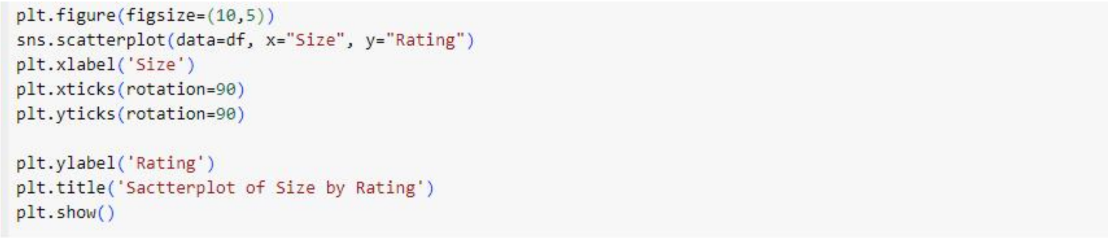
```
missing_values = df.isnull().sum()
print(missing_values)
```

```
App                  0
Category             0
Rating            1474
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
Content Rating       1
Genres               0
Last Updated         0
Current Ver          8
Android Ver          3
dtype: int64
```

We can see a lot of null values.

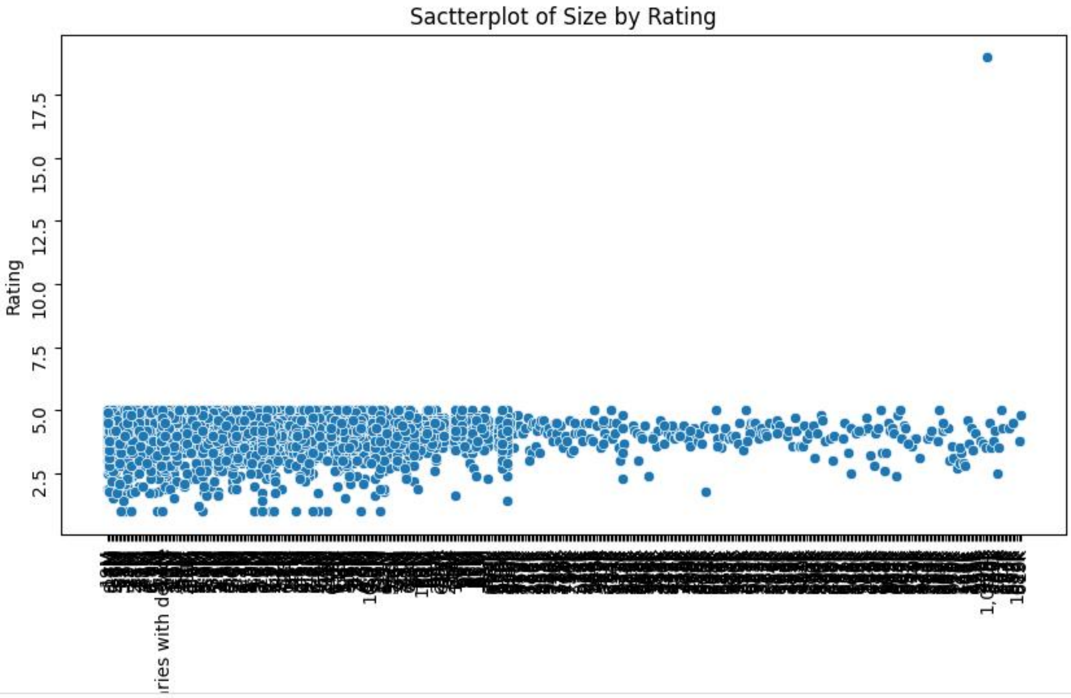
Reduced Sample Size: Missing values reduce the number of observations available for analysis. This reduction in sample size can lead to less reliable statistical estimates and potentially bias your results.
Biased Estimates: If missing values are not randomly distributed across the dataset, they may introduce bias into your analysis. For example, if certain demographic groups are more likely to have missing data, it could skew your findings related to those groups.
Impaired Statistical Power: Missing data can reduce the statistical power of your analysis, making it more difficult to detect significant effects or relationships between variables.
Misleading Results: Ignoring missing values or improperly handling them can lead to misleading results. For instance, if missing values are systematically related to certain variables, omitting them from the analysis could distort the relationship between variables.
Imputation Challenges: Imputing missing values (replacing them with estimated values) can be a solution, but it comes with its own set of challenges. The method used for imputation can affect the results, and imputed values may introduce uncertainty into the analysis.

Question 12



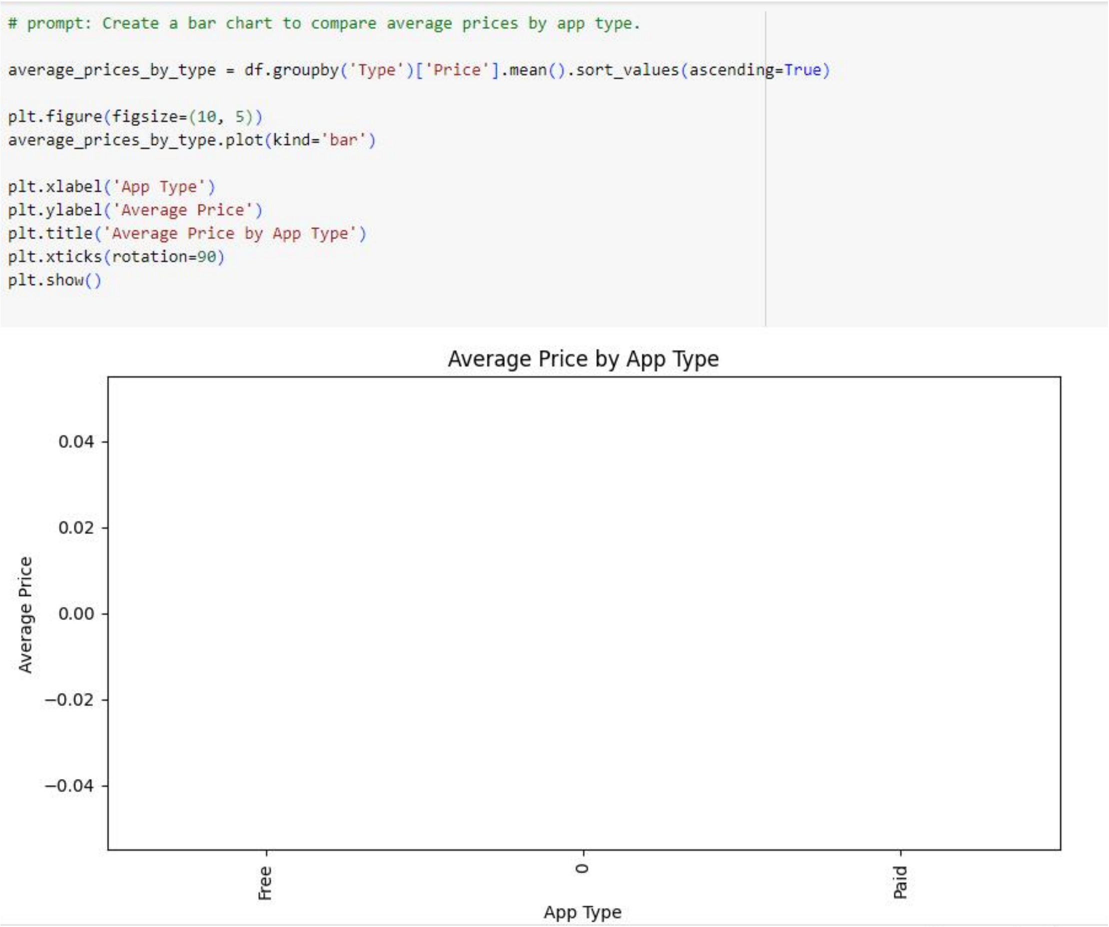
```
plt.figure(figsize=(10,5))
sns.scatterplot(data=df, x="Size", y="Rating")
plt.xlabel('Size')
plt.xticks(rotation=90)
plt.yticks(rotation=90)

plt.ylabel('Rating')
plt.title('Sactterplot of Size by Rating')
plt.show()
```



As the size of the app increases the ratings have decreased.

Question 13



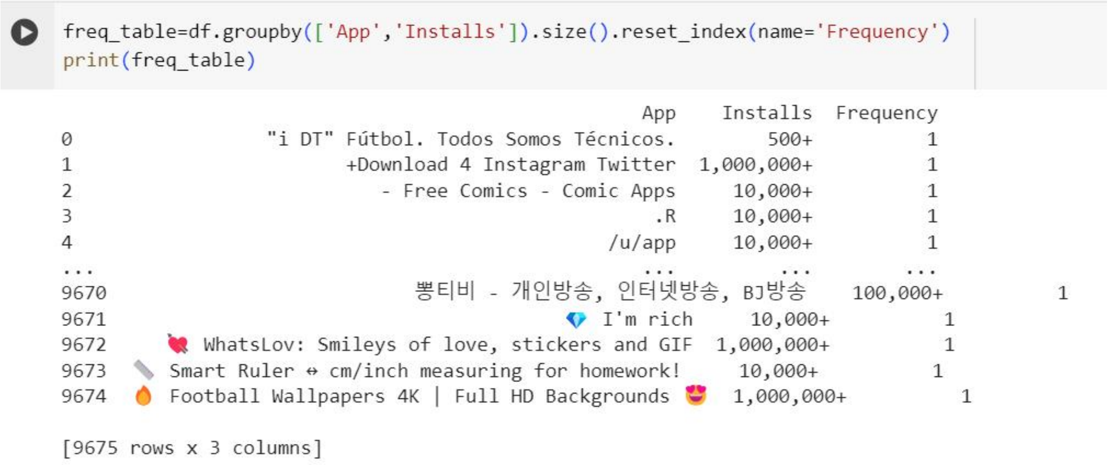
```
# prompt: Create a bar chart to compare average prices by app type.

average_prices_by_type = df.groupby('Type')['Price'].mean().sort_values(ascending=True)

plt.figure(figsize=(10, 5))
average_prices_by_type.plot(kind='bar')

plt.xlabel('App Type')
plt.ylabel('Average Price')
plt.title('Average Price by App Type')
plt.xticks(rotation=90)
plt.show()
```

Question 14



```
freq_table=df.groupby(['App','Installs']).size().reset_index(name='Frequency')
print(freq_table)
```

```
                                               App    Installs  Frequency
0                  "i DT" Fútbol. Todos Somos Técnicos.        500+          1
1                       +Download 4 Instagram Twitter  1,000,000+          1
2                          - Free Comics - Comic Apps     10,000+          1
3                                                  .R     10,000+          1
4                                              /u/app     10,000+          1
...                                               ...         ...        ...
9670                         뽕티비 - 개인방송, 인터넷방송, BJ방송    100,000+          1
9671                                       💎 I'm rich     10,000+          1
9672       💘 WhatsLov: Smileys of love, stickers and GIF  1,000,000+          1
9673  📏 Smart Ruler ↔ cm/inch measuring for homework!     10,000+          1
9674   🔥 Football Wallpapers 4K | Full HD Backgrounds 😍  1,000,000+          1

[9675 rows x 3 columns]
```

Question 15
1.Data Exploration:
Start by loading the dataset and exploring its structure. Understand the columns and the type of data available. Check for any missing values and handle them appropriately, either by removing or imputing them.

2.Data Cleaning:
Clean the data by addressing any inconsistencies, such as misspellings, duplicates, or outliers. Ensure that relevant columns are of the correct data type for analysis.

3.Feature Selection:
Identify the features that are relevant to determining the popularity of an app category. These may include:
App category (e.g., Education, Games, Social)

4. Data Aggregation:
Aggregate the data by app category to calculate metrics such as the average number of installs, average ratings, and total number of reviews for each category.

5.Analysis:
Analyze the aggregated data to identify the most popular app categories based on various metrics.
For example:
Determine the categories with the highest average number of installs or ratings.
Identify categories with a large number of apps but relatively low average ratings or installs, which may indicate a less saturated market.
Consider trends over time by analyzing changes in popularity across different time periods.

6.Visualization:
Visualize the findings using appropriate charts and graphs, such as bar plots, scatter plots, or pie charts. This will help stakeholders easily understand the insights derived from the data.
Recommendations:

Question 16
To identify the most successful app developers in the Google Play Store dataset, I would analyze the following features:

Number of Downloads/Installs:
Apps with a high number of downloads or installs indicate popularity and success.
Ratings and Reviews:
Higher ratings and a large number of positive reviews suggest user satisfaction and engagement with the app.
App Updates:
Regular updates indicate developer commitment to improving the app and addressing user feedback.
App Size:
Smaller app sizes are generally preferred by users, potentially leading to higher adoption rates.
Price:
Free or low-cost apps may attract more users, while premium apps could indicate higher quality or specialized offerings.
Retention Rate:
Apps with high user retention rates are likely to be successful in the long term.
For data visualizations, I would use the following:
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.

Question 17
To determine the best time to launch a new app in the Google Play Store:

- Analyze trend patterns in app installations, ratings, and reviews over time.
- Assess competition and category performance to identify potential niches.

- Utilize line charts for trend analysis, bar charts for comparison, and heatmaps to visualize distribution.
- Evaluate user engagement metrics such as ratings and reviews.
- Employ box plots to understand variability within app categories and make informed launch decisions.

Question
9-17.ipynb