|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ASSIGNMENT COVER SHEET**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | PROGRAMME | : | Master’s in Business Analytics (MsBA) | | | | SUBJECT CODE AND TITLE | : | BAA5033 – Python for Business Analytics | | | | ASSIGNMENT TITLE | : | Python For Business Analytics – Individual Assignment | | | |  |  |  | | | | LECTURER | : | Dr. Aaron Aw Teik Hong | ASSIGNMENT DUE DATE: | 18/10/2024 |   STUDENT’S DECLARATION   1. I hereby declare that this assignment is based on my own work except where acknowledgement of sources is made. 2. I also declare that this work has not been previously submitted or concurrently submitted for any other courses in Sunway University/College or other institutions.   [ Submit “Turn-it-in” report (please tick √): Yes \_\_√\_\_ No \_\_\_\_\_ ]   |  |  |  |  |  | | --- | --- | --- | --- | --- | | NO. | NAME | STUDENT ID NO. | SIGNATURE | DATE | | 1. | Harresh A/L Ragunathan | 19076090 | *Harresh* | 18/10/2024 | |  |  |  |  |  | |  |  |  |  |  | |  |  |  |  |  |   E-mail Address / Addresses (according to the order of names above):   |  |  | | --- | --- | | 1. 19076090@imail.sunway.edu.my |  | |  |  | |  |  |   APPROVAL FOR LATE SUBMISSION OF ASSIGNMENT (If applicable)  IF extension is granted, what is the revised due date? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Signature of Lecturer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   |  | | --- | | Marker’s Comments: |   Marks and / or Grade Awarded: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ADDENDUM**  **USE OF ARTIFICAL INTELLIGENCE (A.I.) DECLARATION**  Students are allowed to use AI to support completion of assessments. However, students are reminded to do so ethically and transparently. This is so that (a) submissions can be fairly and accurately marked; and (b) feedback can be provided on the content that reflects student ability, in order to help with future submissions. Students are also reminded that in accordance with the University’s Academic Malpractice Policy, Item 4.11.2, “*… the representation of work: written, visual, practical or otherwise, of any other person, including another student or* ***anonymous web-based material*** *[emphasis added], or any institution, as the candidate’s own*” is considered malpractice.  **Declaration**  [√] I / We used the following A.I. tools to produce content in this submission:   |  |  |  |  | | --- | --- | --- | --- | | **Tool** | **Purpose** | **Prompts** | **Sections where AI output was used / Outcome(s) in the submission** | | *e.g. ChatGPT* | *e.g. Generating points for the essay*  *Structuring the essay* | *e.g.* “*Give me 5 key talking points for an essay on…”*  *“Show me a structure for an essay on…”* | *e.g. The main point for Section 1.2 and 1.3 were generated by AI, but the discussion was not.*  *The organization / structure of the essay was suggested by AI* | | *e.g. Grammarly* | *e.g. Correcting grammar and spelling, improving sentence structure* | *N/A* | *e.g. Grammarly suggestions were used for all sections of the essay* | |  |  |  |  | |  |  |  |  |   *Note: Add additional rows if necessary.*  **OR**  [ ] I / We did not use any A.I. tools to produce any of the content in this submission.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | NO. | NAME | STUDENT ID NO. | SIGNATURE | DATE | | 1. | Harresh A/L Ragunathan | 19076090 | *Harresh* | 18/10/2024 | |  |  |  |  |  | |  |  |  |  |  | |  |  |  |  |  |   E-mail Address / Addresses (according to the order of names above):   |  |  | | --- | --- | | 1. 19076090@imail.sunway.edu.my |  | |  |  | |  |  | |

# Task 1

|  |
| --- |
|  |

In order to begin the data analysis and visualization, the necessary libraries must first be imported. In this case, the libraries utilized include ‘pandas’, ‘maplotlib’, and ‘seaborn’.

|  |
| --- |
|  |
|  |

The code above loads the data by defining the variable ‘df’ to contain the csv file of the data. By running the ‘df.info()’ code, basic information about each column is outputted in order. This information includes the number of non-null values and what type of data the values are (integer, float, or object). Based on this output, it may be noted that there are 4 categorical variables and 6 numerical variables.

|  |
| --- |
|  |
|  |

The code above outputs the summary statistics for each numerical column. These statistics include the count, mean, standard deviation, minimum, 25th percentile, 50th percentile (median), 75th percentile, and the maximum. Based on the output table, it may be seen that the mean year is approximately 2006, indicating that the majority of the games in this dataset were released in 2006. Furthermore, the minimum and maximum is 1980 and 2020 respectively, indicating the years the oldest and newest games were released in the dataset. It may also be noted that among the four major regions, North America has the highest average sales of 0.265. Looking into the Global Sales, it has an average of 0.534.

|  |
| --- |
|  |
|  |

The code above simply outputs the first five rows of the data.

# Task 2

|  |
| --- |
|  |
|  |

This code outputs both the number of unique genres sold and lists them all. This is done by creating a variable for the unique genres through the use of the “unique.()” function. The number of unique genres is calculated by using the “len.()” function, which returns the number of items in that variable. Based on the output, it can be seen that there are 12 unique genres.

|  |
| --- |
|  |
|  |

This code displays the shape of the data, indicating the number of rows and columns. This is done through the ‘df.shape()’ function, where ‘df.shape[0]’ displays the number of rows, while ‘df.shape[1]’ displays the number of columns. As observed in the output, there are 16,598 rows and 11 columns in this dataset.

|  |
| --- |
|  |
|  |

The term instance is another way of describing a row, which is why ‘df.shape[0]’ is utilized once again. As observed, there are 16,598 instances in this dataset.

|  |
| --- |
|  |
|  |

This code outputs the number of null values in each column through the use of the ‘df.isnull()’ function. It may be observed that the “Year” column has the most missing values (271 missing values), followed by the “Publisher” column (58 missing values).

|  |
| --- |
|  |
|  |

When exploring the original dataset, it may be noticed that the “Global\_Sales” column is a summation of the “NA\_Sales”, “EU\_Sales”, “JP\_Sales”, and “Other\_Sales” columns. This code is made to validate if the values in the “Gloabl\_Sales” column are correct. This is done by creating a ‘calculated\_global\_sales” variable that sums all the regional sales columns. This calculated global sales column will then be compared with the original global sales column using the “!=”operator. By creating a new dataframe while using the “!=” operator, only the values from the calculated sales column that are not equal to the actual global sales column will be stored in the “sales\_discrepancy” variable. It may be noticed that both the “calculated\_global\_sales” and “Global\_Sales” have been rounded to 2 decimal places through the use of the ‘df.round(2)’ function. If these columns were not rounded, it would result in more rows with incorrect sales, which is not accurate. This is the case due to how the “!=” operator functions. For example, numerically, the values “0” and “0.00” are equal. However, the “!=” operator classifies it as not equal. After this conditional data frame has been assigned to a variable, the number of rows with incorrect global sales can be displayed. It may be observed that there are 4,511 rows that are incorrect, which may be cross validated in the csv file for the dataset.

|  |
| --- |
|  |
|  |

After checking the number of incorrect rows, the original “Global\_Sales” column must be corrected. This is done by updating the current “Global\_Sales” data frame to be a summation of the four regional sales, as seen above. The previous code to check for discrepancies may be used again to check if the corrected “Global\_Sales” column has 0 incorrect sales. As seen above, the sales have been corrected.

|  |
| --- |
|  |
|  |

This code displays the top 3 publishers with the highest amount of global sales. This was done by creating a data frame that aggregates the global sales of each publisher. This data frame will then be sorted in descending order, the publisher with the highest sales being displayed first. By using the ‘df.head(3)’ function, the top 3 publishers will be displayed. As observed, the top 3 publishers are Nintendo, Electronic Arts, and Activision.

|  |
| --- |
|  |
|  |

This code is similar to the previous code, except it is related to how many games the top 3 genres have. This is done by creating a data frame that groups the count of games by genre. The top 3 genres are then displayed using the ‘df.head(3)’ function. As seen above, the top 3 genres are Action, Sports, and Misc; Action leading with 3316 games.

# Task 3

|  |
| --- |
|  |
|  |

This code displays a bar chart of the genres based on the count of games. This was done using ‘seaborn’ and ‘matplotlib’. A data frame is created to contain the count of games for each genre in order to plot it into a bar chart. It may be observed that the Action genre has the most games.

|  |
| --- |
|  |
|  |

This code displays a bar chart of the top 5 years with the highest sales. This was done by creating a data frame that aggregates the total global sales for each year and sorts it in descending order based on sales. Then the top 5 years were selected by only selecting the first 5 rows through ‘df.head(5)’. Based on the bar chart, it may be observed that the year 2008 has the most games sold.

|  |
| --- |
|  |
|  |

This code displays a bar chart of the top 5 genres of games sold in Japan. This was done by creating a data frame for the sales in Japan grouped by genre and sorting it in descending order. Based on the bar chart, it may be observed that Role-Playing is the highest earning genre in Japan.

|  |
| --- |
|  |
|  |

The code above displays the count of the top 5 genres. By plotting a bar chart that displays the top 5 genres, the most popular types of games may be identified, which are Action, Sports, Misc, Role-Playing, and Shooter games.

|  |
| --- |
|  |
|  |

Similarly to plotting the top 5 genres, the code above displays the count of the bottom 5 genres played. Instead of using ‘df.head(5)’, ‘df.tail(5)’ was used to display the bottom 5 genres. Based on the bar chart, it may be observed that the least played genres are Puzzle, Strategy, Fighting, Simulation, and Platform games.

|  |
| --- |
|  |
|  |

This code plots a line chart comparing the yearly sales for each region. This was done by creating a dataframe that aggregates all the sales and groups them by year. Based on the line chart, it is noticeable that the Global Sales line slowly increases from 1980, leading to a peak around 2007, followed by a sharp decline approaching 2020. The sales trend lines for each region may also be observed, all reflecting a similar trend with the Global Sales line, but to a lesser degree. However, it should be noted that the Japan’s sales trend line is relatively flat, indicating stable sales throughout the years.

# 

# Task 4

|  |
| --- |
|  |
|  |

When observing the bar chart for the “Top 5 Genres with the Highest Sales in Japan” bar chart from task 3, there are some notable issues regarding its readability and interpretability. First of all, the exact values are hard to identify. This issue was resolved by including gridlines and discrete labels that display the number of sales for each genre’s bar. By including gridlines and labels, viewers will have a much easier experience in finding the exact values of the sale for each genre. Furthermore, in the original bar chart, the color of the bar chart had no value, as each bar was a random color. This issue was resolved by including a color palette that has a gradient, the more intense red indicating high number of sales, while the less intense color indicates a lower number of sales. This allows the viewers to easily identify which genres are selling the most in Japan.

|  |
| --- |
|  |
|  |

An alternative to a bar chart is a pie chart, which visually displays the amount of sales each genre had in Japan based on its area covered in the circle. A pie chart is a suitable alternative because only the top 5 genres are displayed. Since there are only 5 elements, it is easy to identify what the most selling genres are in Japan. Furthermore, the percentages are displayed, indicating the proportion of how much a certain genre is being sold among the 5 genres. Based on the pie chart, among the top 5 genres sold in Japan, 39.8% of the games being sold are within the Role-Playing genre. In the case where more than 5 genres (eg. all genres) are to be displayed, a pie chart would not be suitable, as there will be too many elements of similar values, making it difficult for decision makers to differentiate between genres. Furthermore, there would be too much clutter if labels were included as well. Another alternative, such as a stacked bar chart, would be more suitable as it will be easier to compare the highest earning genres in Japan.

# 

# Task 5

|  |
| --- |
|  |
|  |

This code plots a bar chart displaying the top 10 games sold in North America. This was done by creating a dataframe that sorts the games based on sales in North America in descending order. Then only the top 10 games are plotted by using the ‘df.head(10)’ function. Based on the bar chart, it may be seen that the game that sold the most in North America is “Wii Sports”, followed by “Super Mario Bros.”, and “Duck Hunt”.

|  |
| --- |
|  |
|  |

This code plots a line chart comparing the sales of the four regions throughout the years. This was done by creating a dataframe that aggregates the sales of each region. Based on the line chart, it can be seen that the Japan Sales and Other Sales lines have the most consistent sales compared to North America and Europe. However, it should be further analyzed to determine which of the two regions has the most consistent sales throughout the years.

|  |
| --- |
|  |
|  |

This code plots a bar chart comparing the standard deviations of each region’s sales. Based on the bar chart it can be seen that the variabilities for “JP\_Sales” and “Other\_Sales” are the lowest, aligning with the line chart previously. However, it can now be said that Japan has the most consistent sales among the four regions as it has the lowest standard deviation of 21.23.

|  |
| --- |
|  |
|  |

The code above plots a scatter plot, displaying the correlation between the global sales of the top 10 publishers with the count of games by the top 10 publishers. Based on the output scatter chart, it may be observed that there is a positive but weak correlation between the count of games and sales of games by the top 10 publishers. This may be explained by the trendline, which is relatively flat, indicating a weak correlation.