1. Find a data that you would like to use. This dataset should include at least 3 numerical variables and 2 categorical variables. You are allowed to use or collect your own data. Preprocess the data before the analysis. Explain how, where you collected/retrieved your sample (Data).

I have collected the data from Kaggle.com which is largest hub for dataset for the research of Data Analyst. And I

Have selected top 100 Ranking Video game sales from 16500 video games. Where following column are there.

Name: Name of game

Platform: Release platform for the game

Year: Year of release

Genre: Game genre

Publisher: Company which released the game

NA\_Sales: North American sales

EU\_Sales: European sales

JP\_Sales: Japan sales

Other\_Sales: Sales in other countries

Global\_Sales: Sales around the worlds

1. Employ descriptive and frequency analysis and report the analysis results to a formatted table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table.1 Descriptive Statistics | | | | | | | |
|  | N | Minimum | Maximum | Mean | Std. Deviation | Skewness | Kurtosis |
| NA\_Sales | 100 | 0.98 | 41.49 | 7.0499 | 5.873226 | 3.267 | 14.057 |
| EU\_Sales | 100 | 0.01 | 29.02 | 4.2066 | 3.511557 | 4.007 | 24.761 |
| JP\_Sales | 100 | 0 | 10.22 | 2.0462 | 2.048607 | 1.155 | 1.552 |
| Other\_Sales | 100 | 0.08 | 10.57 | 1.3177 | 1.559055 | 3.849 | 17.906 |
| Global\_Sales | 100 | 7.34 | 82.74 | 14.6198 | 9.936065 | 3.821 | 21.893 |
| Valid N (listwise) | 100 |  |  |  |  |  |  |

*Note: -The sales of the video games are in million dollars.*

Descriptive Analysis has been performed with the sample of 100 data from top 100 ranking sales of video game throughout the worldwide. (Skewness=3.82 and Kurtosis-21.83)

Sales around the word has most sales apart from the other continent with 7.34 is the minimum where as 82.74 million for Maximum sales. However, an Average sale is 14.61 million has been done.

Similarly, In North America, 0.98 million sales are the second minimum and 41.49 is the maximum sales with an Average of 7.04 million. (Skewness=3.26 and Kurtosis-14.04)

And in Japan, some video games have not been sold with 10.22 million maximum sales along 1.31 million least sales among the worldwide sales. (Skewness=3.84 and Kurtosis-17.906)

|  |  |  |
| --- | --- | --- |
|  | Frequency | Percent |
| Activision | 14 | 14 |
| Atari | 1 | 1 |
| Bethesda Soft works | 1 | 1 |
| Electronic Arts | 5 | 5 |
| Microsoft Game Studios | 6 | 6 |
| Nintendo | 52 | 52 |
| Sega | 1 | 1 |
| Sony Computer Entertainment | 8 | 8 |
| Square Soft | 1 | 1 |
| Take-Two Interactive | 9 | 9 |
| Ubisoft | 2 | 2 |
| Total | 100 | 100 |

On the Basis of given 100 sample data, Frequency Analysis has been done. Where, Nintendo publisher is the one most producing company which has published 52 video games. And second most video games produced by Activision which is 14 video games.

1. Employ correlation analysis.

To Employee the correlation analysis, I would like to perform the sales of video game between North America and Europe sales of video games.

**Hypothesis Stated:**

H0: ρ(rho) =0

(The variable NA sales and Europe Sales are independent. There is no significant relationship between X and Y)

HA: ρ(rho) ≠ 0

(The variable NA sales and Europe Sales are dependent. There is a significant relationship between X and Y)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table:3 Tests of Normality | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| NA\_Sales | 0.221 | 100 | 0.000 | 0.679 | 100 | 0.000 |
| EU\_Sales | 0.194 | 100 | 0.000 | 0.673 | 100 | 0.000 |

On the basis of 100 sample data collected, Anderson Darling of Normality Test has been performed. According to Kolmogrov-smirnov, P-Value is less than alpha and According to Shapiro-Wilk test, P-value is also less than our alpha as well. But I will follow, Shapiro test because, data collection has less than 2000. Since, p-values is less than 5%. Which I have enough evidence to reject the null hypothesis. Thus, I can conclude that sales Data and Europe sales data are not normally distributed.

We can employ Spearman’s Rank Order Correlations analysis since at least one of the variables is non-normally distributed.

|  |  |  |  |
| --- | --- | --- | --- |
| Table:4 Correlations | | | |
|  |  |  | NA\_Sales |
| Spearman's rho | EU\_Sales | Correlation Coefficient | 310\*\* |
|  |  | Sig. (2-tailed) | 0.002 |
|  |  | N | 100 |
| \*\* Correlation is significant at the 0.01 level (2-tailed). | | | |

On the basis of given data collection, Spearman’s Rank Order Correlations analysis, there is a positive, moderate, and statistically significant relationship between NA\_Sales and EU\_Sales (r=31%, p-value=.002<0.05). I investigated the Correlation-ship analysis among the North America and Europe sales, Normality test was performed for the data distribution. After Normality test was performed, Spearman’s Rank was performed. Since p-Value is less than alpha, we can conclude that we have enough evidence to reject the null hypothesis in favor of alternative hypothesis. Hence, Sales of NA and Sales of EU are dependent.

1. Propose a hypothesis that might employ cross table with the chi-square test.

On the given collection of top 100 ranking sample data from of sales of video games throughout the world wide, I have investigated the cross table of chi-square test between publisher and Average Sales where sales are categories into binary number i.e. 0 and 1 where O is considered as Sales of publisher which is less than average of total sales and “1” is considered as sales of publisher greater than Average.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table:5. Publisher \* Average Cross-tabulation | | | | | |
|  |  |  | Average | | Total |
|  |  |  | Less Average  (14.6198 ) | Greater Average  (14.6198) | |
| Publisher | Activision | Count | 12 | 2 | 14 |
|  |  | % within Publisher | 85.70% | 14.30% | 100.00% |
|  |  | % within Average | 17.60% | 6.30% | 14.00% |
|  | Atari | Count | 1 | 0 | 1 |
|  |  | % within Publisher | 100.00% | 0.00% | 100.00% |
|  |  | % within Average | 1.50% | 0.00% | 1.00% |
|  | Bethesda Soft works | Count | 1 | 0 | 1 |
|  |  | % within Publisher | 100.00% | 0.00% | 100.00% |
|  |  | % within Average | 1.50% | 0.00% | 1.00% |
|  | Electronic Arts | Count | 5 | 0 | 5 |
|  |  | % within Publisher | 100.00% | 0.00% | 100.00% |
|  |  | % within Average | 7.40% | 0.00% | 5.00% |
|  | Microsoft Game Studios | Count | 5 | 1 | 6 |
|  |  | % within Publisher | 83.30% | 16.70% | 100.00% |
|  |  | % within Average | 7.40% | 3.10% | 6.00% |
|  | Nintendo | Count | 28 | 24 | 52 |
|  |  | % within Publisher | 53.80% | 46.20% | 100.00% |
|  |  | % within Average | 41.20% | 75.00% | 52.00% |
|  | Sega | Count | 1 | 0 | 1 |
|  |  | % within Publisher | 100.00% | 0.00% | 100.00% |
|  |  | % within Average | 1.50% | 0.00% | 1.00% |
|  | Sony Computer Entertainment | Count | 7 | 1 | 8 |
|  |  | % within Publisher | 87.50% | 12.50% | 100.00% |
|  |  | % within Average | 10.30% | 3.10% | 8.00% |
|  | Square-Soft | Count | 1 | 0 | 1 |
|  |  | % within Publisher | 100.00% | 0.00% | 100.00% |
|  |  | % within Average | 1.50% | 0.00% | 1.00% |
|  | Take-Two Interactive | Count | 5 | 4 | 9 |
|  |  | % within Publisher | 55.60% | 44.40% | 100.00% |
|  |  | % within Average | 7.40% | 12.50% | 9.00% |
|  | Ubisoft | Count | 2 | 0 | 2 |
|  |  | % within Publisher | 100.00% | 0.00% | 100.00% |
|  |  | % within Average | 2.90% | 0.00% | 2.00% |
| Total |  | Count | 68 | 32 | 100 |
|  |  | % within Publisher | 68.00% | 32.00% | 100.00% |
|  |  | % within Average | 100.00% | 100.00% | 100.00% |

From the given Chi-Square analysis (Table:5), From the top 100 ranking data provided, 68% of the sales from all publisher is found to be less than average sales were 32% is found to be greater than Average sales. However, Nintendo video games are biggest number of game producer with 52 number of game where 53% of Nintendo published games are below average and rest is above average sales.

Secondly, Activision publisher is the second most sales with 14 numbers of video games with 85.70% sales are below Average and 14.30% are above average sales.

Thirdly, Take-two Interactive is third most publishing game with 9 number of games along 55.6% sale are below average and 44.4% sales are above average.

Moreover, there are Atari, Bethesda Soft works, Sega and Square-soft publisher who has only games published where their all sales are below average done in the whole world.

1. Propose a hypothesis that might employ a parametric or non-parametric test to investigate the comparison of a variable between groups (It can be two groups or three or more groups).

H0 = µ1- µ2- µ3 =0

H1 = µ1- µ2 -µ3 ≠0

Where µ1 is the mean value of sales of Nintendo publisher

Where µ2 is the mean value of sales of Activision publisher.

Where µ3 is the mean value of sales of Take-Two Interactive publisher.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table:6 Tests of Normality | |  |  |  |  |  |  |
|  | Publisher | Kolmogorov-Smirnova | | | Shapiro-Wilk | |  |
|  |  | Statistic | df | Sig. | Statistic | df | Sig. |
| Global\_Sales | Activision | 0.23 | 14 | 0.043 | 0.877 | 14 | 0.053 |
| Nintendo | 0.208 | 52 | 0 | 0.712 | 52 | 0 |
| Take-Two Interactive | 0.161 | 9 | .200\* | 0.939 | 9 | 0.575 |

According to Normality test based on given data, Shapiro-Wilk test has 0.05 which is equal to p-value and based on Kolmogorov-Simonov Test, the significance value is 0.043 which is less than 5%. We can say that data for Game published by Activision is not normally distribution.

Similarly, based on the Shapiro-Wilk test of normality, p-value is 0 which is less than alpha. Hence, it is also not normally distribution.

Thus, we can conclude that we have enough evidence to reject the null hypothesis in support of Alternative hypothesis.

And, Data are not normally distributed. And we don’t need to prove the normality test for another group because, if one

Group is found to be not normally distributed. Hence we can Employee non-parametric of Kruskal-wallis Test.

|  |  |  |
| --- | --- | --- |
| Table:- 7 Ranks |  |  |
| Publisher | N | Mean Rank |
| Activision | 14 | 30.64 |
| Nintendo | 52 | 39.69 |
| Take two Interaction | 9 | 39.67 |
| Total | 75 |  |

On the basis of three group, the result has been sown in table 7. With total number of 75 data and mean ranking

are 30.64 for Activision, 39.69 for Nintendo and 39.67 for Take two interactions.

|  |  |
| --- | --- |
| Table : -8 Test Statisticsa,b | |
|  | Global Sales |
| Chi-Square | 1.962 |
| df | 2 |
| Asymp. Sig. | .375 |

As Kruskal-wallis analysis done between three group which are Activision, Nintendo and Take two-Interaction, it has been found that, significance value is more than our alpha which is 37.5% > 5%. Which we don’t have enough evidence to reject the null hypothesis and can conclude that, the mean value of three groups are equal.

1. Propose a multiple regression model that will investigate the impact of the proposed independent variables on the proposed dependent variables. Investigate this proposed model thoroughly.

The aim of this study is to investigate is impact of Platform (categorical), Genre, and Year, Global sales, to find the impact on global sales average. In order to test these associations a multiple regression analysis was utilized. The sample, included TOP 100 ranking video games sales from all over country. The regression analysis is shown in table 12.

**Proposed model**

**Global Sales = b0 + b1\*X1 + b2\*X2 + b3\*X3 + b4\*X5 +b5\*X6+b6\*X7 + e**

**Where b0, b1, b2, and b3 are samples coefficient.**

And

X1 = platform

X2 =Genre

X3 =EU

X4 =Other Sales

X5=JP Sales

X6 = NA sales

The VIF values of the independent variables ranged from 1.078 to 1.974 which is far from lower than cutoff value 10. Therefore, there was no multi-collinearity issue. The results indicated that the platform, Genre, Europe Sales, Japan Sales. North America sales and other sales could reliable predicted the Global sales. Thus, the overall model significant (F=412691 and P-value < 5% from Table No: -9 & 10)

**Table: - 9 Coefficients**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Independent Variable | | Coefficients | t | Sig. | VIF |
| 1 | (Constant) | -.037 | | -.780 | .437 |  |
|  | Platform | -4.224E-05 | | -.320 | .750 | 1.659 |
|  | Genre | .000 | | .875 | .384 | 1.078 |
|  | EU\_Sales | 1.000 | | 3957.485 | .000 | 1.974 |
|  | Other\_Sales | 1.001 | | 2031.919 | .000 | 1.478 |
|  | JP\_Sales | 1.000 | | 2385.169 | .000 | 1.851 |
|  | NA\_Sales | 1.000 | | 7373.727 | .000 | 1.591 |

**Table No:10 -Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model |  | Df | F | Sig. |
|  | Regression | 6 | 41269189.52 | 0.000 |
|  | Residual | 93 |  |  |
|  | Total | 99 |  |  |

Moreover, 46.6% of the variance in Global Sales by genre, Platform, Europe, North America, Japan and other sales (Adj R2=48.8% from table: -11.

**Table: -11 Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | 1.000a | 1.000 | 1.000 | .006283 |

The results revealed that Genre value has not relation at all nor significant effect on global sales (b=-0.00, P-value=38.4%).

Similarly, Platform has negative and not significance effects on global sales with b=-4.224E-005 and p-value=75%. Which is greater than 5%.

Similarly, EU sales has positive and significantly effects on global sales with b=1.000 and p-value=0.00%. Which is less than 5%.

Similarly, JP sales has positive and significantly effects on global sales with b=1.000 and p-value=0.00%. Which is less than 5%.

Similarly, NA sales has positive and significantly effects on global sales with b=1.000 and p-value=0.00%. Which is less than 5%.

And EU sales has positive and significantly effects on global sales with b=1.000 and p-value=0.00%. Which is less than 5%.

Table:12 Multiple Regression

|  |  |
| --- | --- |
| Independent variable | Model  Dependent Variable: Global Sales |
| Genre | 0.00\*\*\*  (38.4%) |
| Platform | 1\*\*\*  (75%) |
| EU sales | 1\*\*\*  (0.00) |
| NA sales | 1\*\*  (0.00) |
| Japan Sales | 1.00\* |
|  | (0.00) |
| Other sales | 0.32\*\*\*  (8.96) |
| F-Stat | 41269189.52\*\*\* |
| Adj-R2 | 100% |
| N | 100 |

\*P<0.05; \*\*p<0.01; \*\*\*p<0.001

1. Propose a logistic regression model that will investigate the impact of the proposed independent variables on the proposed dependent variables. Investigate this proposed model thoroughly.

On the basis of given data, the effect of sales of North America, Japan, Europe and other continent on Global sales was

investigated using binary logistic regression analysis. The Analysis results were shown in Table.

The samples included top 100 ranking of Video games sold in all over the world. The dependent variable

Bin\_Average\_Globalsales was coded as 1 if the global sales is above average sales and “0” if sales is below. All the step

taken to analysis the effect by using logistic regression is shown below figure 1.

From the data collection for the Logistic regression, 98.04% is the correct and 1.96% is incorrect

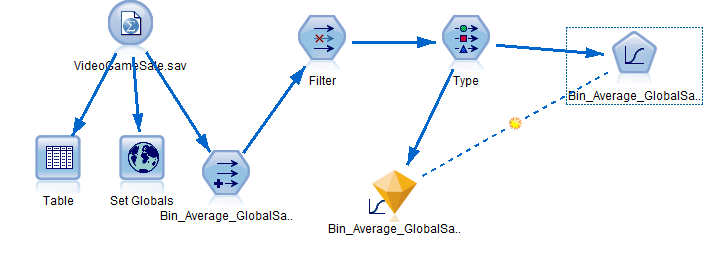


Figure . Step taken to determine Logistic Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table:13 Model | Model Fitting Criteria |  | | |
|  | -2 Log Likelihood | Likelihood Ratio Tests | | |
|  |  | Chi-Square | df | Sig. |
| Intercept Only | 125.374 |  |  |  |
| Final | 0 | 125.374 | 4 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| Table:14 Analysis for Binary\_Average-Globalsales | | | |
| Correct | 100 | 98.04% |  |
| Wrong | 2 | 1.96% |  |
| Total | 102 |  |  |

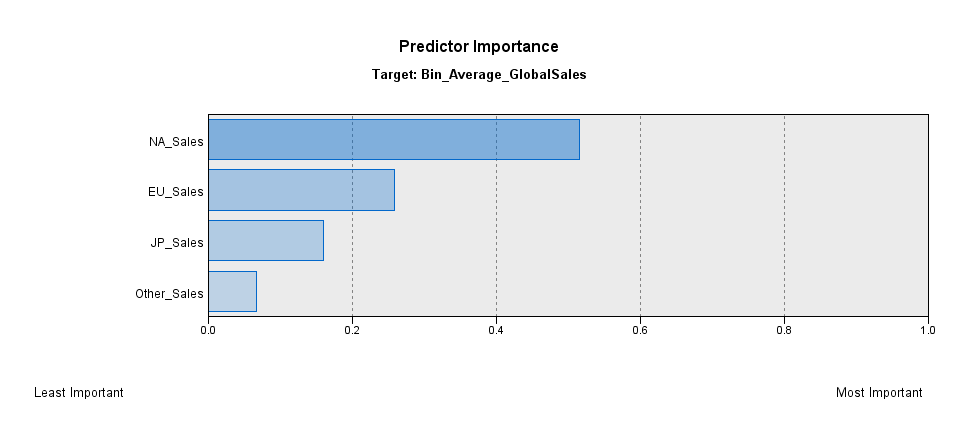


Figure . Predictor Importance

**Table: -15 Predictor importance data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nodes | Importance | Importance | V4 | V5 |
| Other\_Sales | 0.0669 | 0.0669 | Other\_Sales | 0.0669 |
| JP\_Sales | 0.1594 | 0.1594 | JP\_Sales | 0.1594 |
| EU\_Sales | 0.2579 | 0.2579 | EU\_Sales | 0.2579 |
| NA\_Sales | 0.5157 | 0.5157 | NA\_Sales | 0.5157 |

Table: -16 Parameter Estimation

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bin\_Average\_GlobalSalesa | B | Std. Error | Wald | df | | | Sig. | | Exp(B) | |
| Intercept | -534.683 | 15526.02 | 0.001 | | 1 | 0.973 | | |  | | |
| NA\_Sales | 38.442 | 1176.886 | 0.001 | | 1 | 0.974 | | | 4.957E+16 | | |
| EU\_Sales | 41.856 | 1251.813 | 0.001 | | 1 | 0.973 | | | 1.506E+16 | | |
| JP\_Sales | 32.235 | 921.898 | 0.001 | | 1 | 0.972 | | | 9.990E+16 | | |
| Other\_Sales | 17.644 | 2287.896 | 0 | | 1 | 0.994 | | | 45984713.59E+16 | | |
| a The reference category is: 0. | | |  | |  |  | |  | |  | | |  |
| b Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing. | | | | | | | | | | | | | |

**Proposed model**

**Bin\_Average\_Globalsales = b0 + b1\*X1 + b2\*X2 + b3\*X3 + b4\*X4 + e**

**Where b0, b1, b2, and b3 are samples coefficient.**

And

X1 =EU

X2 = NA sales

X3=JP Sales

X4 = Other Sales

North America sales with 51.57% is the most important factor, so as Europe sales with 25.79% is second most important factor. Similarly, Japan with 15.94% is third most important factor on Global sales where as other sales with 6% is the least one important factor that affects the global sales. (From Table:15 and figure: 2)

The odd ratio was used in order to determine the effect estimate of the variables. The results indicated that Europe sales, North America sales, Japan sales and other sales are reliable predict the global sales. (-2LR stat= 125.374, Chi-Square=125.374 and p-Value <5%).

The result showed that for a one point increased of North America sales, the odd ratio for Global sales is increased by 38.442 and p-value is 97.4% < 5% which means increasing in global sales is not significantly associated with an increased likelihood of global sales.

Similar, the result showed that for a one point increased of Japan sales, the odd ratio for Global sales is increased by 32.235 and p-value is 97.2% < 5% which means increasing in global sales is not significantly associated with an increased likelihood of global sales.

Similar, the result showed that for a one point increased of Europe sales, the odd ratio for Global sales is increased by 41.856 and p-value is 97.3% < 5% which means increasing in global sales is not significantly associated with an increased likelihood of global sales.

And, the result showed that for a one point increased of other sales, the odd ratio for Global sales is increased by 17.644 and p-value is 99.4% < 5% which means increasing in global sales is not significantly associated with an increased likelihood of global sales.

Table: -16 Logistic Analysis for Bin\_Average\_Globalsales dependent variable.

|  |  |
| --- | --- |
| Independent variable | Model  Dependent Variable: Global Sales |
| EU sales | 41.856\*  (4.957E+16) |
| NA sales | 38.442\*\*  (1.506E+18) |
| Japan Sales | 32.235\*  (9.990E+13) |
| Other sales | 17.644\*\*\*  (45984713.59) |
| -2LogLikeliHood | 125.374 |
| Likelihood Ratio Test(x2/df) | 125.374 |
| N | 100 |
| Nagelkerke approximation of R2 |  |

1. After analyzing the data using the analysis tools, report the results using formatted tables and interpret the analysis results in a single write-up file.
2. Upload a) the data file b) all the analysis outputs, c) and the **TERM PROJECT WRITE UP**.

|  |  |
| --- | --- |
| **Criteria** | **Points** |
| The Write- UP includes all the necessary interpretations of the particular analysis including the data collection process. Appropriate explanations and/or comments regarding the results are included. | .../15 |
| The tables are formatted, cleaned, and well presented | .../10 |
| The required analysis tools are employed | .../5 |
| [http://turnitin.com/](#_1fob9te). The project will be subjected to Turnitin.  The overall similarity rate is less than 20%. The writing of the analysis results is original.  Those with more than 50% will be penalized. Please check the ACADEMIC HONESTY POLICY on the syllabus. | .../5 |
| The data includes the necessary variables for the analysis | .../5 |
| Total | .../40 |

**Deadline: December 3, 2018; 11:59pm**